The American Economic Review

ARTICLES 3

MARCH 1977

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President
LAWRENCE R. KLEIN
University of Pennsylvania

President-Elect
JACOB MARSCHAK
University of California, Los Angeles

Vice Presidents
ROBERT EISNER
Northwestern University

ANNE O. KRUEGER
University of Minnesota

Secretary
C. ELTON HINSHAW
Vanderbilt University

Treasurer and Editor of the Proceedings
RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review
GEORGE H. BORTS
Brown University

Managing Editor of The Journal of Economic Literature
MARK PERLMAN
University of Pittsburgh

Executive Committee

Elected Members of the Executive Committee
CAROLYN SHAW BELL
Wellesley College
BURTON A. WEISBROD
University of Wisconsin, Madison
EDMUND S. PHELPS
Columbia University
ALICE M. RIVLIN
Congressional Budget Office
ROBERT J. LAMPMAN
University of Wisconsin, Madison
MARC NERLOVE
Northwestern University

Ex Officio Members
ROBERT AARON GORDON
University of California, Berkeley
FRANCO MODIGLIANI
Massachusetts Institute of Technology

• Published at George Banta Co., Inc., Menasha, Wisconsin.

• THE AMERICAN ECONOMIC REVIEW, including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and Supplements, is published by the American Economic Association and is sent to all members five times a year, in February, March, June, September, and December.

Association membership dues for 1977, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows

$25.00 for regular members with rank of assistant professor or lower, or with annual income of $12,000 or less.

$30.00 for regular members with rank of associate professor, or with annual income of $12,000 to $20,000,

$35.00 for regular members with rank of full professor, or with annual income above $20,000;

$12.50 for junior members (registered students) Certification must be submitted yearly

Subscriptions (libraries, institutions, or firms) are $37.50 a year Only subscriptions to both publications will be accepted Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $3.70 to cover extra postage.

• Correspondence relating to the *Papers and Proceedings*, the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C Elton Hinshaw, 1313 21st Avenue South, Nashville, Tennessee 37212. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices Printed in U.S.A.

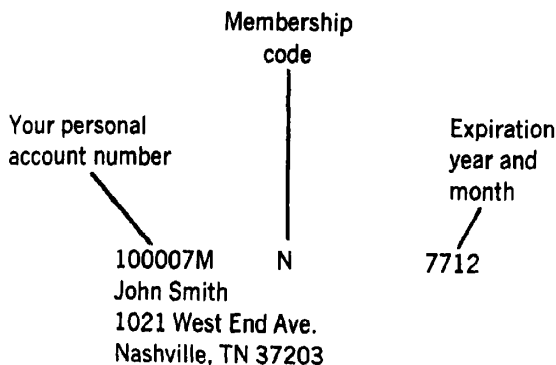
CONTENTS OF ARTICLES AND SHORTER PAPERS

CONTENTS OF PAPERS AND PROCEEDINGS

CONTRIBUTORS TO ARTICLES AND SHORTER PAPERS

CONTRIBUTORS TO PAPERS AND PROCEEDINGS

AMERICAN ECONOMIC ASSOCIATION COMPUTERIZES

In October 1977, the Association's mailing list was converted to a computerized system. It is now EXTREMELY IMPORTANT that you send your address label with all of your correspondence — whether it be changes of address, payments, complaints, etc. If you are unable to send the actual Address Label, please include your Account Number (which is described below).

Membership Codes

REGULAR MEMBERS (code "M") with the academic rank of Assistant Professor or lower or with annual income of $12,600 or less irrespective of rank, paying an annual fee of $26.25.

REGULAR MEMBERS (code "N") with the academic rank of Associate Professor or with annual income above $12,600 but not more than $21,000, paying an annual fee of $31.50.

REGULAR MEMBERS (code "O") with the academic rank of Full Professor or with annual income above $21,000, paying an annual fee of $36.75.

JUNIOR MEMBERS (code "J"), available to registered students for three years only, paying an annual fee of $13.00. Student status must be certified by your major professor or school registrar.

FAMILY MEMBERS (code "F"), available for two or more persons living at the same address — receiving only one set of publications of the Association. Second membership pays an annual membership fee of $5.25.

We realize that there may be some problems in the beginning, but we are confident that the operation of the Association will be much more efficient with the computerization of the lists.

Concepts of Optimality and Their Uses

By TJALLING C. KOOPMANS*

According to a frequently cited definition, economics is the study of "best use of scarce resources." The definition is incomplete. "Second best" use of resources, and outright wasteful uses, have equal claim to attention. They are the other side of the coin.

For our present purpose the phrase "best use of scarce resources" will suffice. However, each of the two nouns and two adjectives in this phrase needs further definition. These definitions in turn need to be varied and adjusted to fit the specific circumstances in which the various kinds of optimizing economic decisions are to be taken.

I will assume that the main interest of this gathering is in the range of applications of the idea of best use of scarce resources, and in the ways in which the main categories of applications differ from each other. I shall therefore describe mathematical ideas and techniques only to an extent helpful for the exploration of that range of applications.

A good place to start is with the *production programs of the individual plant or enterprise* for a short period ahead. The "resources" then include the capacities of the various available pieces of equipment. In a centrally directed economy they may also include the allotments of nationally allocated primary inputs such as fuels, raw materials, labor services. In a market economy with some capital rationing one single allotment of working capital available for the purchase of primary inputs at given market prices would take the place of most of the primary input allotments.

In either institutional framework, an especially simple prototype problem is obtained if one fixes the quantity of required output of the product or products made by the enterprise, while prices for the primary products are given. Then the term "use" of resources stands for a choice of a technical process or a combination of processes that meets that requirement within the given constraints. "Best" use is a, or if unique, that choice that meets the requirement at minimum cost of primary inputs.

Economists have differed as to whether this problem belongs in economics. In the 1920's the British economist A. C. Pigou stated "... it is not the business of economists to teach woolen manufacturers how to make and sell wool, or brewers how to make and sell beer...."

This was not the attitude of economists in several other European countries. In particular, there was in the 1930's a lively discussion among Scandinavian and German economists concerning models of production possibilities and their use in achieving efficiency within the enterprise. The Nordisk Tidsskrift for Teknisk Økonomi provided an important medium for these discussions. Significant contributions[1] were made by Carlson, Frisch, Gloerfelt Tarp, Schmidt, Schneider, Stackelberg, and Zeuthen.

Thus the situation at the end of the 1930's was one in which important practical problems in the best use of resources within the enterprise had been neglected by economists in several countries, and had been taken up by only a handful of economists in a few other countries. In addition, the problems were of a kind in which special knowledge possessed by other professions, mathematicians, engineers, managers, was pertinent. One could therefore have expected important new contributions to come from these neighboring professions.

This is precisely what happened, and several times over. Chronologically first was the publication by the mathematician

*Yale University. This article is the lecture Tjalling Koopmans delivered in Stockholm, Sweden, December 1975, when he received the Nobel Prize in Economic Science. and is included in the volume of *Les Prix Nobel en 1975*

[1]For references to these authors and to quotation, see the author (1957, p. 185).

Leonid V. Kantorovich (1939, in Russian) of a 68-page booklet entitled, in translation (1960), "Mathematical Methods of Organizing and Planning of Production." The importance of this publication is due to the simultaneous presence of several ideas or elements, some of which had also been present in earlier writings in different parts of economics or mathematics. I enumerate the elements.

(1) *A model of production* in terms of a finite number of distinct production processes, each characterized by constant ratios between the inputs and outputs specific to the process.

This element has a long history in economics. It is found in Walras (1874, Leçon 41; 1954, Lesson 20), Cassel (1919, ch. 4), the mathematician von Neumann (1936), Leontief (1936, 1941), all dealing with models of the productive system as a whole. However, the feature most important for our purpose was present only in the classical writers in the theory of international trade[2] and in the models of von Neumann and of Kantorovich. This feature is that the output of one-and-the-same required commodity can in general be achieved by more than one process. The same specified vector of outputs of all required commodities can therefore in general be obtained as the outcome of many different combinations of processes. Two such combinations may differ in the list of processes included—and in the levels of activity assigned to the processes they both use. It is due to this element of choice between alternative ways of achieving the same end result that a genuine optimization problem arises. It is true that Walras also optimized (1954, Lesson 36) on the choice of processes, but from an infinite collection defined by a differentiable production function. It is precisely this choice of a more general collection of processes that delayed the recognition by economists of the applications that are our present topic:

(2) The perception of a wide range of *practical applications* of the model to industries that themselves are sources for the data required by these applications.

These included the transportation problem to be discussed below, an agricultural problem, and various industrial applications. The definition and collection of available data of a different, more aggregative, kind was also an important element in Leontief's input-output analysis.

(3) The demonstration that with an optimal solution of the given problem, whether of cost minimization or output maximization, one can associate what in Western literature has been called *shadow prices,* one for each resource, intermediate commodity or end-product.

Kantorovich's term in 1939 was "resolving multipliers," which he changed to "objectively determined valuations"[3] in his book of 1959. In general, these valuations are equal to the first derivatives of the negative of the cost minimum, with respect to the specified availabilities of the goods in question. In mathematical terminology these valuations have also been called "dual variables," in contrast with the activity levels assigned to the processes, which are then called "primal variables." Analogous dual variables occur also in von Neumann's model of proportional growth, with an interpretation as prices in competitive markets.

(4) The identification of a *separation theorem* for convex sets due to Minkowski as a mathematical basis for the existence of the dual variables.[4]

(5) The *computation* of optimal values of the primal and associated dual variables for illustrative examples, and some indi-

[2]See the references to Torrens (1815), Ricardo (1817), Mill (1852), Graham (1923), and others in the survey article by John Chipman.

[3]*Ob'ektivno obuslovlennye otsenki.*

[4]For this purpose von Neumann had used the heavier tool of a topological fixed-point theorem. The dispensibility of this for his purpose was shown later by David Gale (1956), and by Koopmans and Bausch (Topic 5).

cations toward calculating such solutions in more complex cases.

Finally, brief but precise explanations of

(6) The interpretation of the dual variables as defining equivalence ratios (*rates of substitution*) between different primary inputs and/or different required outputs, and

(7) The additional interpretation of the dual variables as *guides for* the coordination of *allocative decisions* made in different departments or organizations.

I shall return to (7) below.

Kantorovich's work of 1939 did not become known in the West until the late 1950's or early 1960's. Meanwhile the transportation model was redeveloped in the West without knowledge of the work on this topic by Kantorovich (1942, reprinted 1958) and Kantorovich and Gavurin (1940, 1949). The Western contributions were made by Hitchcock (1941), the author (memo dated 1942, published 1970; articles of 1949 and 1951 (with Reiter)), Dantzig (ch. 23 in Koopmans, ed. 1951).

The general linear model was rediscovered and developed by George Dantzig and others associated with him, under the initial stimulus of the scheduling problems of the U.S. Air Force. The term "linear programming" came into use for the mathematical analysis and computational procedures associated with this model. A compact early publication of this work can be found in a volume entitled *Activity Analysis of Production and Allocation*, edited by the author. Substantial further developments appeared in such journals as *Econometrica*, *Management Science*, *Operations Research*, and were brought together in Dantzig's *Linear Programming and Extensions* (1963), a book that was many years in the making. These developments, in which many mathematicians and economists took part, went substantially beyond the earlier work of Kantorovich, in several directions. I note only a few of the extensions to the elements listed above.

(2′) *Extension of the range of applications* to animal feeding problems, inventory and warehousing problems, oil refinery operations, electric power investments[5] and many other problems.

(3′), (4′) Further clarification of the *mathematical relations between primal and dual variables* and their extension to *nonlinear programming* by[6] Tucker, Gale, Kuhn and others.

This work also traced additional mathematical origins or precursors for the duality theory of linear programming in the work on game theory by von Neumann (1928 and, with Morgenstern, 1944) and by Ville (1938), and in work on linear inequalities by Gordan (1873), Farkas (1902), Stiemke (1915), Motzkin (1936) and others.[7]

(5′) The development by Dantzig of the *simplex method* for maximizing a linear function under linear contraints (including inequalities) and the further improvements to this method by Dantzig and others.

The simplex method has become the principal starting point for a family of algorithms dealing with linear and convex nonlinear allocation problems. These methods can be set up so as to compute optimal values of both primal and dual variables.

Most important to economic theory as well as application was a further extension of (7) into

(7′) The analysis of the *role or use of prices* toward best allocation of resources, either through the operation of competitive markets, or as an instrument of national planning.

These ideas, again, have a long history in economics. In regard to competitive markets, they go back at least as far as Adam Smith (1776), and were eloquently restated and developed by Hayek (1945). Important writers on the use of prices in socialist planning were Barone (1908), Lange (1936), and

[5]See, for instance, Massé and Gibrat.

[6]See Gale, Kuhn, and Tucker; Kuhn and Tucker; and, for a summary, Tucker.

[7]For references, see Dantzig, chs. 2 3.

Lerner (1937, 1938). The new element in the work by the author (1949, 1951) and Samuelson (1949, 1966) was the use of the linear model and, in my own case, the attempt to develop what may be called a *pre*-institutional theory of allocation of resources. It was already foreshadowed in the work of Lange and Lerner that hypothetical perfect competition and hypothetical perfect planning both imply efficient allocation of resources—although neither occurs in reality. It therefore seemed useful to turn the problem around, and just postulate allocative efficiency as a model for abstract, pre-institutional study. Thereafter, one can go on to explore alternative institutional arrangements for approximating that model.

I believe that the linear model offers a good foothold for this purpose. First, it makes a rigorous discussion easier. Secondly, the most challenging *non*linearity—that connected with increasing returns to scale—in fact undermines competition. It also greatly escalates the mathematical and computational requirements for good planning. The linear model, therefore, makes a natural first chapter in the theory of best allocation of resources. In its simplest form it leads to the following symmetric relationships between activity levels of the processes and the (shadow) prices of the resources and goods produced:

(7″) (a) Every process in use makes a zero profit,
(b) No process in the technology makes a positive profit,
(c) Every good used below the limit of its availability has a zero price,
(d) No good has a negative price.

These same relationships are a recurrent theme in the first two chapters of Kantorovich's (1959) book, which also was many years in the making prior to publication. It was subsequently translated into French and English, the latter under the title *The Best Use of Economic Resources*. The gist of the book's recommendations is that socialist planning can achieve best attainment of the goal set by the planning body through calculations that ensure the fulfillment of these or similar conditions for optimality.

Kantorovich did not go much beyond his earlier remarks on the questions concerning possible use of a price system for decentralization of decisions. This became a major theme, however, in the abstract work of Koopmans (1951, Sec. 5.12), in the work on two-level planning by Kornai and Lipták (1965) in relation to planning in Hungary, and in that by Malinvaud (1967) stimulated by experiences with planning in France. The principal computational counterpart of this work was developed by Dantzig and Wolfe (1960, 1961) under the name "the decomposition principle."

The third chapter of Kantorovich's book deals with the problem of investment planning to enlarge the production base. The principal emphasis is on the concept of the *normal effectiveness of capital investment*. This is a discount rate applied to future returns and to contemplated investments and other future costs, in the evaluation and selection of investment projects. This idea had been proposed earlier by Novozhilov (1939). The point emphasized by Kantorovich is that the prices to be used in calculating returns and costs should be the objectively determined valuations determined by his methods, for the selection to have an optimal result. These proposals were at the time new to the practice of Soviet economic planning. I believe that the principle of the normal effectiveness of capital investment has gained increasing acceptance in Soviet theory and practice since that time. There is an obvious formal analogy with the profitability criterion for investment planning used by the firm in a market economy, using anticipated market prices and the appropriate market rate of interest. The institutional framework contemplated is, of course, fundamentally different. I believe that the underlying pre-institutional optimizing theory is the same.

Summing up, I see two principal merits in the developments I have reviewed so far. One is their initially pre-institutional character. Technology and human needs are universal. To start with just these elements has facilitated and intensified professional contacts and interactions between economists from market and socialist countries.

The other merit is the combination and merging of economic theory, mathematical modeling, data collection, and computational methods and algorithms made possible by the modern computer. A genuine amalgam of different professional contributions!

The linear model, followed by the convex nonlinear model, have provided the proving ground for these developments—and—their most conspicuous limitation: The nonconvex nonlinearities associated with increasing returns to scale, i.e., with the greater productivity of large-scale production in many industries, require quite different methods of analysis, and also raise different problems of institutional frameworks conducive to best allocation.

I now proceed to a rather different class of applications of the idea of best allocation of scarce resources. This field is usually referred to as the *theory of optimal economic growth*. In most studies of this kind made in the countries with market economies there is not an identifiable client to whom the findings are submitted as policy recommendations. Nor is there an obvious choice of objective function, such as cost minimization or profit maximization in the studies addressed to individual enterprises. The field has more of a speculative character. The models studied usually contain only a ew highly aggregated variables. One then onsiders alternative objective functions hat incorporate or emphasize various trands of ethical, political, or social hought. These objectives are then tried out o see what future paths of the economy hey imply under equally simplified assumpions of technology or resource availability. he principal customers aimed for are other conomists or members of other profesions, who are somewhat closer to the makng of policy recommendations. These may e those engaged in making more disaggreated optimizing models of growth that inorporate numerical estimates of technoogical or behavioral parameters. (I shall turn to this field of "development proamming" below.) Or the hoped-for cusmers may be policy economists who may find it useful to have the more abstract ideas of this field in the back of their minds when coping with the day-to-day pressures for outcomes rather than criteria.

The question of the clientele is even more baffling when the problem concerns growth paths for time spans covering several generations. What can at best be recommended in that case is the signal the present generation gives, the tradition it seeks to strengthen or establish, for succeeding generations to take off from.

The classic in the optimal growth field is a paper published in 1928 by Frank Ramsey, known also as the author of equally fundamental papers on the foundations of mathematics and on subjective probability. His definition of "best" involves the maximization of a sum (or integral) of utility flows to be derived from future consumption. Using a continuous time variable, Ramsey's choice of objective function is a limiting case of a broader class of functions which I shall consider first,

$$U = \int_0^\infty e^{-\rho t} u(c_t)\, dt, \qquad 0 < \rho,$$

the *objective function representing generations*. Here c_t denotes the aggregate consumption flow as of time t, and $u(c)$ is a utility flow serving as an evaluating score for the consumption flow c. One chooses the function $u(c)$ so as to increase with c but at a decreasing rate du/dc as c increases. This expresses that at all times "more is better," but less so if much is already being enjoyed (see Figure 1). The effect on the allocation of consumption goods between generations is similar to the effect of a progressive income tax on spendable incomes among contemporaries.

We shall call the exponentially decaying factor $e^{-\rho t}$, $\rho > 0$, *the discount factor for utility*. It diminishes the weight given to future utility flows in the summation of the entire utility flow over all the future to form a total score U. The weight is smaller the larger the *discount rate for utility* ρ, and the further one looks into the future. On ethical grounds Ramsey would have none of this. I shall take the view that the important question of discounting utility—or for that mat-

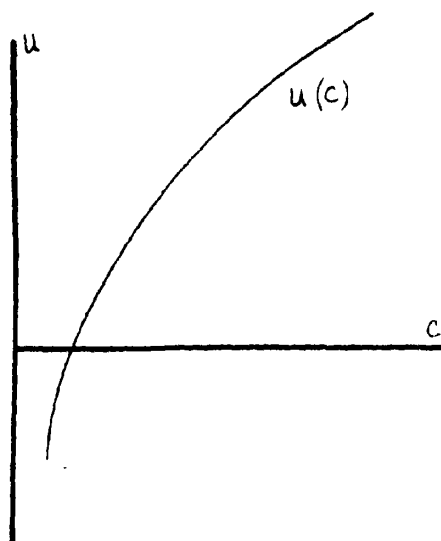
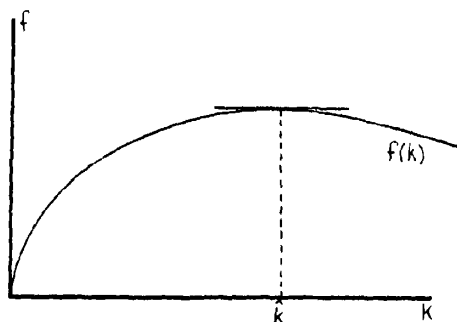



FIGURE 1

FIGURE 2

ter any other aspect of the choice of the objective function—should not be settled entirely on a priori grounds. Most decision makers will first want to know what a given objective function will make them do in given circumstances. I shall therefore hold $\rho > 0$ for this first exploration,[8] and turn to the mathematical modeling of the "circumstances" in terms of technological and resource constraints on the consumption and capital variables.

One "resource" is the labor force. It need not enter the formulae because it will be assumed to remain inexorably constant over time. The only other resource is an initial capital stock denoted k_0, historically given as of time $t = 0$. The "use" at any time t of labor and of the then capital stock k_t consists of two steps. The first and obvious step is to achieve at all times the highest net output flow $f(k_t)$ that can be produced by the labor force, using the capital stock fully and to best effect. The form given to the function $f(k)$ summarizes and simplifies broad technological experience. It specifies $f(0) = 0$ ("without capital no output"), $f(k)$ initially increasing with k but at a diminishing rate df/dk, in such a way that from some point $\hat{k}$ of capital saturation on, $f(k)$ decreases because depreciation rises more steeply than gross output. (See Figure 2.)

In all this the product flow $f(k)$ is regarded as consisting of a single good, which can be used as desired for consumption or for adding to the capital stock,

$$f(k_t) = c_t + \frac{dk_t}{dt},$$

the *allocation constraint*. To determine this allocation for all t is the second step. This is done "best" at all future points of time if the total score U is thereby maximized.

It might seem as if this constrained maximization problem is quite different in mathematical structure from those discussed before. This is not the case. The main difference is that the discussion has shifted from a vector space to a function space, using conventional notations not designed to reveal the common structure of the two problems. In particular, as long as the crucial convexity assumptions are maintained, interpretations in terms of shadow prices remain valid.

The problem for $\rho > 0$ was solved independently by Cass (1966), Koopmans (1965, 1967), Malinvaud (1965, 1967), thirty-five years after Ramsey. Without proof I indicate the nature of the solution in Figure 3. In the diagram on the left, the *abscissa k* is set out along the *vertical* half-axis, the

[8]For an objective function implying a variable discount rate that depends on the path contemplated, see Koopmans (1960); Koopmans, Diamond, and Williamson; and Beals and Koopmans.

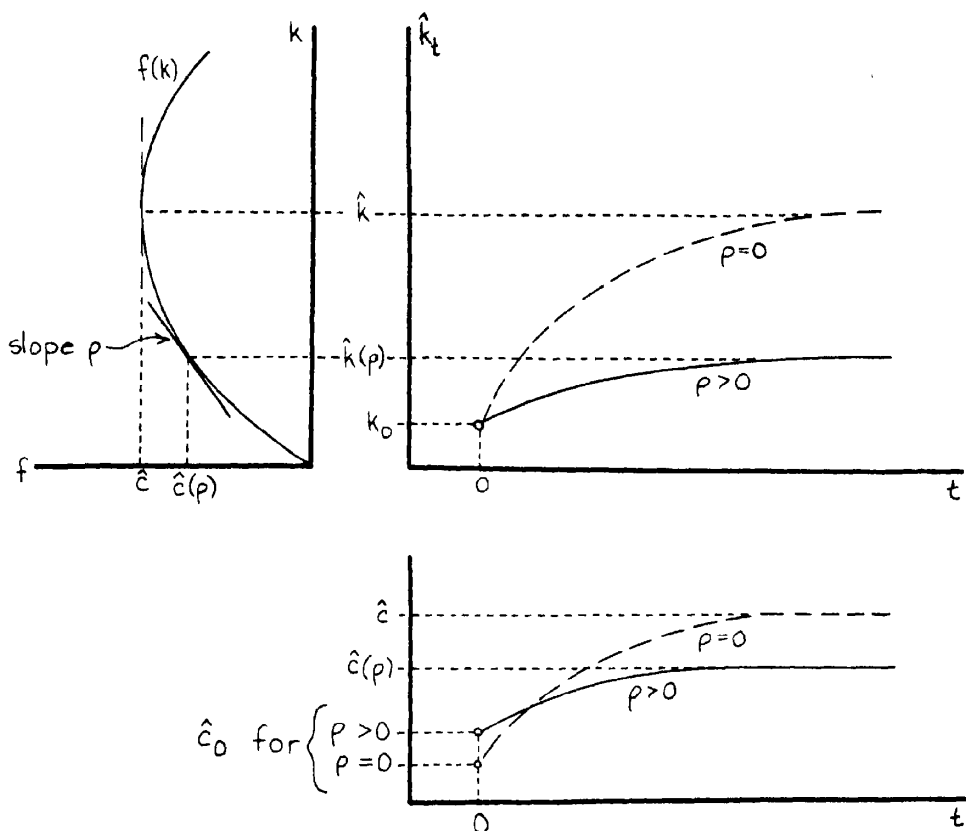



FIGURE 3

ordinate $y = f(k)$ along the *horizontal* half-axis pointing left. For given $\rho > 0$, find the unique point $\hat{k}(\rho)$ on the curve $y = f(k)$ in which the slope df/dk equals ρ. Then, if the initial capital stock k_0 should happen to equal $\hat{k}(\rho)$, the optimal capital path remains constant, $k_t = \hat{k}(\rho)$, over all the future. For any initial stock k_0 less than or larger than $\hat{k}(\rho)$, the optimal path shows a monotonic and asymptotic approach to $\hat{k}(\rho)$. All this is illustrated in solid lines in the two top diagrams in Figure 3. The lower right diagram shows the corresponding optimal consumption path c_t, which approaches the asymptotic level $\hat{c}(\rho) = f(\hat{k}(\rho))$.

The differential equation which, together with the allocation constraint given above, governs the approaches of k_t and c_t to their asymptotes is subject to an interesting interpretation in terms of shadow prices. Define such prices, p_t for the consumption good and q_t for the use of a unit of the capital stock, by

$$p_t = e^{-\rho t}\left(\frac{du}{dc}\right)_{c=c_t}$$

$$q_t = p_t\left(\frac{df}{dk}\right)_{k=k_t}$$

Then p_t is the marginal utility of consumption du/dc, at the level c_t of consumption flow reached at time t, discounted back to time zero. In turn, q_t is the marginal productivity of capital df/dk, at the level k_t reached at time t, multiplied by the shadow price p_t of the product, already defined. This makes q_t the marginal productivity of capital in terms of discounted utility.

In these terms the differential equation is

$$q_t = -\frac{dp_t}{dt}$$

the *allocation principle*. If p_t and q_t were to be market prices for the capital good and for its use in production, this equality would state that the returns on two alternative dispositions for the capital good would be equal: One is to sell the good now, the other is to sell first only its use for a short period, and thereafter to sell the good itself. That this principle is necessary for optimality is intuitively plausible. Its sufficiency can be proved from the convexity assumptions about the functions $-u(\cdot)$ and $-f(\cdot)$, plus the boundary condition that

$$\lim_{t\to\infty} k_t = \hat{k}(\rho)$$

What is the effect of choosing different values of the discount rate ρ? Figure 3 suggests the answer for the realistic case that the initial capital stock k_0 is well below its ultimate level $\hat{k}(\rho)$. In that case, as ρ is *de*creased, that is, as the present valuation of consumption in a distant future is *in*creased, then the asymptotic levels of the capital stock and of consumption are both *in*creased. However, to accumulate the additional capital that makes this feasible, consumption in the present and the near future is further *de*creased. Thus, the impatience expressed by a positive discount rate merely denies to uncounted distant generations a permanently higher level of consumption because that would necessitate a substantially smaller present consumption. Perhaps a pity, but not a sin.

Ramsey showed that the effect of a decrease in ρ goes right down to but not beyond the limiting case of no discounting, $\rho = 0$. The optimal paths for that case are shown by dashed lines in Figure 3. Ramsey used an ingenious mathematical device that gets around the nonconvergence of the utility integral for $\rho = 0$, and also leads to a proof simpler than the one for positive ρ.

This narrow escape for virtue is blocked off by some quite plausible modifications of the model toward greater realism. For instance, one may introduce a population (= labor force) L_t that changes with time and, for interpersonal equity, modify the objective function to read

$$V = \int_0^\infty e^{-\rho t} L_t u(c_t) dt$$

the *objective function representing individuals*, where c_t is now interpreted as *per capita* consumption, k_t as ditto capital. The thought here is to give equal weight (apart from discounting) to all future individuals rather than generations. Assume further that L_t cannot be influenced but is subject to an exogenous exponential growth

$$L_t = L_0 e^{\lambda t}, \qquad \lambda > 0$$

over time. By way of example, let λ correspond to a growth by 3 percent per year. Then, for mathematical reasons alone, the discount rate ρ has to correspond to at least 3 percent per year for an optimal path to exist. If one tries to keep ρ at zero and force existence by imposing a *finite* time horizon of one century, say, then the "optimization" produces an irrational and arbitrary pileup of consumption toward the end of that century while the capital stock runs down to zero—or to any other prescribed level that still leaves room for a terminal splurge.

The model discussed so far leaves out important aspects of the modern economy. I shall comment briefly on the incorporation into the model of 1) exhaustible resources; 2) technological change; 3) population as a policy variable.

With regard to *exhaustible resources*, I shall only consider the extreme case of an *essential* resource. By this I mean a resource that is essential to sustaining life, is not capable of complete recycling, and has no substitute either now or later within the remaining period of its availability.

I have constructed a highly simplified model (1973) of an essential resource which leads to conclusions quite different from those reached for the preceding *capital model*. The only process in the technology consists of costless extraction of the resource combined with its immediate and direct consumption. This process is avail-

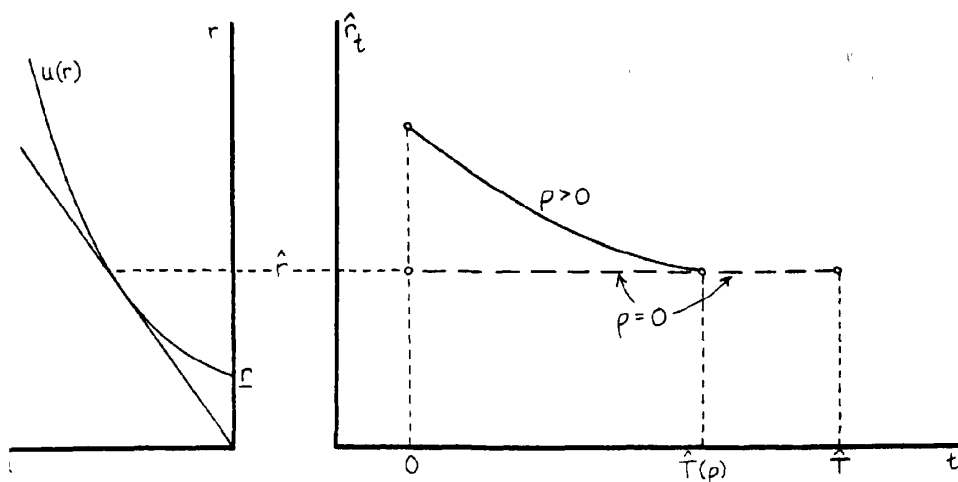



FIGURE 4

ible at all times until the time T at which he resource is exhausted. Population is assumed given and constant for the *survival period* $0 \leqq t \leqq T$. At time T it falls to zero. In these circumstances, higher per capita consumption by those living early enough o share in the available resource shortens he survival period, hence reduces the total number sharing. Thus the survival period is now a policy variable, and thereby so is population in some part of the future.

We now adopt again the objective function of the Ramsey model with discounting, except that the integration extends only over the period of survival. Then, unlike in he capital model, the optimal path now depends on where one places the zero point in he utility scale in which the function $u(\cdot)$ is expressed. Let us set this zero point somewhat arbitrarily at the utility flow level, $u(\underline{r}) = 0$, of the resource consumption level below which life cannot be sustained. The resulting function $u(r)$ is shown (again sideways) on the left in Figure 4.

In the absence of discounting ($\rho = 0$), ime now enters only as a scale on which the fatal cutoff point T is determined. Among hose so admitted equal sharing is clearly optimal, hence optimal consumption stays at a constant level over time. Figure 4 shows dashed lines) how this level, and therewith he survival period, are determined: All included claimants consume optimally at that unique per capita level $\hat{r}$ that maximizes the utility flow per unit of resource flow consumed. Note that this consumption level is well above the subsistence minimum $\underline{r}$. This result is to be expected from an objective function that places value not on the number of included people as such, but only on the "number of utils" enjoyed by all alive taken together.

Discounting ($\rho > 0$) now introduces unequal weights, decreasing as time goes on, between people with equal technological opportunities. The result is (solid line) that the consumption of earlier claimants is raised over that of later ones, with the last claimants no better off than before. Of necessity, the cutoff point arrives sooner, and some of those included for $\rho = 0$ will not be there to press their claim if $\rho > 0$.

I hold no brief for the realism of this model. I have brought it up only as a stark demonstration of the point that the problem of whether and how much to discount future utilities cannot be equitably resolved a priori and in the abstract. One needs to take into account the opportunities expected to be available to the various consumers now and later, for the given technology and resource base.

Does there exist an essential resource, as defined above, exhaustible in less than

astronomical or even geological time? I have argued elsewhere (1973) that the best available answer to this question must come from those natural scientists and engineers most able to assess what future technology may do, or be made to do, to find substitute materials or fuels for those now within sight of exhaustion. I should add that geologists may well develop ways to pin down further the best estimates of ultimate availability. Economists could bring up the rear with methods for integrating the diverse pieces of information so obtained.

This leads to a few remarks on models that recognize *technological change*, and the uncertainty by which it is inevitably surrounded. A conceptual step forward in this direction has been made by Dasgupta and Heal (1974). Their model postulates an exhaustible resource for which a substitute will or may become available at an uncertain future date. The uncertainty is described by a subjectively estimated probability distribution. Capital accumulation is also represented in the model. The objective function is the mean value of the distribution of the sum of discounted utilities over an infinite future period. The optimal path to be found consists of two successive segments. One is the path to be followed from $t = 0$ up to the as yet unknown time $t = T$ of the availability of the substitute. The other is the path to be followed from time T on, which of course depends on the situation in which the advent of the substitute finds the economy.

In one interesting sub-case, the effect of uncertainty can be represented by an equivalent addition to the discount rate—until the substitute is available. This is the case in which the technological change is expected to be so incisive that the old capital stock will lose all its value as a result of the availability of the new technology.

Deep problems of choice of optimality criterion arise in models in which *population size is a decision variable*, or is affected by decision variables. I shall contrast only two distinct criteria. One, used in a population context by Meade (1955), Dasgupta (1969), and others, is given by the "objective function representing individuals" already discussed. This function multiplies the utility of per capita consumption by the number of consumers before forming the discounted sum. The other criterion, the "function representing generations," sums just the per capita utilities.

The argument by Arrow and Kurz (1970, I.4) in favor of the former criterion uses the analogy of two contemporaneous island populations under one government. It shows convincingly that per capita utilities for the two islands must before their addition be multiplied by the respective populations if one wants to avoid discrimination between people of the two groups.

Does this argument carry over to generations that succeed each other in time? One might well argue this on the same grounds if population increase is a truly exogenous function of time, unalterable by any policy, but fortunately one that permits a feasible path. One such case was already considered above. But in fact population *can* be influenced, directly by persuasion and provision of information, or indirectly through other economic variables. I submit that this makes a difference.

A distinction should be made here between concern for the welfare of those already alive and of those as yet unborn whose numbers are still undecided. Current social ethic urges recognition of the needs and desires of living persons, within nations, and between nations—even though practice differs from norm. But there is something open-ended about the same concern for our descendants. How many descendants?

The answer to that question is different for the two criteria we are now comparing. On the basis of indications in a recent paper of Pitchford (1975, p. 21) I conjecture[9] that the criterion representing individuals will, under constraints like those considered above, recommend a smaller per capita

[9]For this conjecture to be well defined, the criterion must again specify the zero point in the utility scale. I set it again at the subsistence minimum, because associating a zero utility with any higher consumption level would imply a preference for the phasing out of human life in some adverse circumstances in which a very low consumption level permitting survival remains feasible.

ıtility for all generations, present and uture, than does the criterion representing ;enerations. This is to be expected because he criterion representing individuals at- aches value to numbers as such. Specifi- :ally it would rate the combination of a ;iven per capita utility level and population ize below another combination with a 5 ›ercent lower utility, say, and a more than ı percent larger population.

So the issue appears to be one of quantity 'ersus quality. Again, one would want to ry out the two criteria, and other ones, ınder a variety of constraints before a fully :onsidered judgment can be made. But the simple idea of adding up individual utilities does not seem to me compelling in itself. Those already born are committed to their existence by the instinct of self-preserva- tion. Choosing a criterion that limits births so as to allow a good existence to an in- definitely continuing sequence of genera- tions appears sufficient as an end in itself. Why more people at the expense of each of them?

The foregoing discussion tacitly assumes that net population increase can be con- trolled without much delay, within the range recommended by optimization. In reality, neither the techniques nor the acceptance of birth control are such as to support that assumption. On the other hand, it is not so that population increase can be regarded as entirely exogenous. This suggests refine- ment of the models in two directions.

First, a more realistic model should in- corporate estimates of the relations describ- ing the response of reproductive behavior to levels of income, education, housing, med- ical care, and other causative variables. In :ircumstances where the resulting path of population does not seriously reduce per :apita income below what could be achieved by a more direct population policy, no further action would be required. Where this is not the case, the processes whereby reproductive behavior can be influenced, and in particular the relation between re- source inputs into these processes and the responses to them (prompt and delayed) need to be incorporated in the model. Con- siderations of this kind have been intro- duced into optimal population models by Pitchford (1974).

The second refinement concerns the optimality criterion. Situations occur in which what can be achieved by population policies can only diminish but not prevent a lowering of the per capita consumption sustained by domestic resources, for an ex- tended period ahead. In such cases a realistic view will recognize an exogenous com- ponent in the future path of population for some time to come. One may then want to explore optimality criteria that extend to those as yet unborn children whose birth had better, but cannot, be prevented the same consideration as to those already born.

We have considered two broad fields of application for optimization models. One comprises the detailed and data-oriented optimization of the decisions of the enter- prise or public agency – and also the coor- dination of such decisions through a price system, through centralized planning and management, or both. The other is the more speculative study of alternative aggregate future growth paths for an entire economy.

In conclusion I want to make some re- marks about the growing field of *develop- ment programming*, in which the two strands of thought are being combined and merged. One early step in this development was the construction of a mathematical program- ming model for an economy as of some future year, including investments and the flow of aid in the intervening period as de- cision variables. An example is the study of the economy of southern Italy by Chenery, writing with Kretschmer and with Uzawa. An evaluative description of experiences with Hungarian economic planning along these lines was written by Kornai. Later studies, such as that of the Mexican econ- omy by Goreux, Manne et al. envisaged a sequence of future years. In most of these studies data availabilities determined the use of Leontief's input-output framework for representing the production possibilities of the economy as a whole. Policy choices and optimization were introduced where data so permitted. One example is the

choice between domestic production versus imports paid for by exports in the southern Italy study. Others are the sectoral detail in the Hungarian studies, and concentration on the energy and especially electric energy sectors in the Mexican one.

A weakness in the treatment of consumption in optimal growth models, noted by Chakravarty, is the lack of continuity between consumption levels in the past and those recommended by otherwise reasonable looking optimality criteria for the near future. One remedy proposed by Manne (1970) has been to constrain future consumption paths to a family of smooth paths all anchored on the most recent observed level of consumption.

Econometric studies have been used to estimate consumption, production and investment relations describing decisions not or only partly controlled by the policy maker. In some cases the studies have gone beyond the convexity assumptions of most optimal growth models to recognize economies of scale in production. Examples are studies by Chenery (1952) of investment in pipelines in the United States, and by Manne et al. (1967) of plant size, location and timing of availability in four industries in India.

A substantial part of the work in this field may have escaped the general reader of economic journals, because much of the work has been published in collective volumes. Examples of these, not already mentioned, are Manne and Markowitz, Adelman and Thorbecke, Chenery, Blitzer et al.

One final remark. The economist as such does not advocate criteria of optimality. He may invent them. He will discuss their pros and cons, sometimes before but preferably after trying out their implications. He may also draw attention to situations where allover objectives, such as productive efficiency, *can* be served in a decentralized manner by particularized criteria, such as profit maximization. But the ultimate choice is made, usually only implicitly and not always consistently, by the procedures of decision making inherent in the institutions, laws, and customs of society. A wide range of professional competences enters into the preparation and deliberation of these decisions. To the extent that the economist takes part in this decisive phase, he does so in a double role, as economist, and as a citizen of his polity: local polity, national polity, or world polity.

REFERENCES

Irma Adelman and Erik Thorbecke, *The Theory and Design of Economic Development,* Baltimore 1966.

Kenneth Arrow and Mordecai Kurz, *Public Investment, the Rate of Return, and Optimal Fiscal Policy,* Baltimore 1970.

E. Barone, "Il Ministro della Produzione nello Stato Collettivista," *Giorn. degli Econ.*, 1908, translation in Friederich A. Hayek, ed., *Collectivist Economic Planning,* London 1935.

R. Beals and T. C. Koopmans, "Maximizing Stationary Utility in a Constant Technology," *SIAM J. Appl. Math.*, Sept. 1969, *17*, 1001–15.

Charles R. Blitzer et al., *Economy-Wide Models and Development Planning,* Oxford 1975.

D. Cass, "Optimum Growth in an Aggregate Model of Capital Accumulation: A Turnpike Theorem," *Econometrica,* Oct. 1966, *34*, 833–50.

Gustav Cassel, *The Theory of Social Economy,* London 1923.

Sukhamoy Chakravarty, *Capital and Development Planning,* Cambridge, Mass. 1969.

Hollis B. Chenery, *Studies in Development Planning,* Cambridge, Mass. 1971.

——— "Overcapacity and the Acceleration Principle," *Econometrica,* Jan. 1952, *20*, 1–28.

——— **and K. Kretschmer,** "Resource Allocation for Economic Development," *Econometrica,* Oct. 1956, *24*, 365–99.

——— **and H. Uzawa,** "Non-Linear Programming in Economic Development," in Kenneth Arrow et al., eds., *Studies in Linear and Non-Linear Programming,* Stanford 1958.

J. S. Chipman, "A Survey of the Theory of International Trade: Part I, The Classical Theory," *Econometrica,* July 1965, *33*, 477–519.

eorge B. **Dantzig**, *Linear Programming and Extensions*, Princeton 1963.
——, "Application of the Simplex Method to a Transportation Problem," in Tjalling Koopmans, ed., *Activity Analysis of Production and Allocation*, New York 1951.
—— **and P. Wolfe**, "The Decomposition Principle for Linear Programming," *Oper. Res.*, Jan. 1960, *8*, 101–11.
—— **and** ——, "The Decomposition Algorithm for Linear Programs," *Econometrica*, Oct. 1961, *29*, 767–78.
. S. **Dasgupta**, "On the Concept of Optimum Population," *Rev. Econ. Stud.*, July 1969, *36*, 295–318.
—— **and G. Heal**, "The Optimal Depletion of Exhaustible Resources," *Rev. Econ Stud., Symposium*, 1974, 3–28.
). **Gale**, "The Closed Linear Model of Production," in Harold Kuhn and A. W. Tucker, eds., *Linear Inequalities and Related Systems*, Princeton 1956.
——, **H. W. Kuhn, and A. W. Tucker**, "Linear Programming and the Theory of Games," in Tjalling Koopmans, ed., *Activity Analysis of Production and Allocation*, New York 1951.
. **M. Goreux, A. S. Manne et al.**, *Multi-Level Planning: Case Studies in Mexico*, Amsterdam 1973.
'. **A. Hayek**, "The Use of Knowledge in Society," *Amer. Econ. Rev.*, Sept. 1945, *35*, 519–30.
'. **L. Hitchcock**, "The Distribution of a Product from Several Sources to Numerous Localities," *J. Math. and Physics*, Apr. 1941, *20*, 224–30.
.eonid **V. Kantorovich**, *Mathematischeskie Metody Organizatsii i Planirovania Proizvodstva*, Leningrad 1939; translated as "Mathematical Methods in the Organization and Planning of Production," *Manage. Sci.*, July 1960, *6*, 366–422.
—— **and M. K. Gavurin**, "Primenenie matematicheskikh metodov v voprosakh analiza gruzopotokov," in *Problemy povysheniia effektivnosti raboty transporta* ("The Use of Mathematical Methods in Analyzing Problems of Goods Transport," in *Problems of Increasing the Efficiency in the Transport Industry*, 110–38), Academy of Sciences, USSR 1949.
——, *Ekonomicheskii raschët nailichshego ispolzovania resursov*, Acad. of Sc., USSR 1959; translated as *The Best Use of Economic Resources*, Cambridge, Mass. 1965.
——, "On the Translocation of Masses," *Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS*, 1942, *37*, nos. 7–8; reprinted in *Manage. Sci.*, Oct. 1958, *5*, 1–4.
Tjalling C. Koopmans, *Activity Analysis of Production and Allocation*, New York 1951.
——, *Three Essays on the State of Economic Science*, New York 1957.
——, "Exchange Ratios between Cargoes on Various Routes," (memo dated 1942), in *Scientific Papers of Tjalling C. Koopmans*, Berlin 1970.
——, "Optimum Utilization of the Transportation System," in *Proc. Internat. Statist. Conferences*, 1951, *5*, 136–45.
——, "Analysis of Production as an Efficient Combination of Activities," in his *Activity Analysis of Production and Allocation*, New York 1951.
——, "Stationary Ordinal Utility and Impatience," *Econometrica*, Apr. 1960, *28*, 287–309.
——, "On the Concept of Optimal Economic Growth," in *The Econometric Approach to Development Planning*, Amsterdam 1965.
——, "Intertemporal Distribution and 'Optimal' Aggregate Economic Growth," in William Fellner et al., eds., *Ten Economic Studies in the Tradition of Irving Fisher*, New York 1967, 95–126.
——, "Economic Growth and Exhaustible Resources," in H. C. Bos et al., eds., *Economic Structure and Development, Essays in Honour of Jan Tinbergen*, Amsterdam 1973, 239–55.
—— **and S. Reiter**, "A Model of Transportation," in *Activity Analysis of Production and Allocation*, New York 1951.
—— **and A. F. Bausch**, "Selected Topics in Economics Involving Mathematical Reasoning," *SIAM Rev.*, July 1959, *1*, 79–148.

——, **P. A. Diamond, and R. E. Williamson,** "Stationary Utility and Time Perspective," *Econometrica*, Jan. 1964, *32*, 82-100.

János Kornai, *Mathematical Planning of Structural Decisions*, Budapest 1967.

—— **and Th. Liptak,** "Two-Level Planning," *Econometrica*, Jan. 1965, *33*, 141-69.

H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," in J. Neyman, ed., *Proc. Second Berkeley Symposium on Math. Statist. and Probability*, Berkeley 1950, 481 92.

O. Lange, "On the Economic Theory of Socialism," *Rev. Econ. Stud.*, Oct. 1936, *4*, 53 71; Feb. 1937, *4*, 123-42; reprinted in Benjamin Lippincott, ed., *On the Economic Theory of Socialism*, Minneapolis 1938.

W. W. Leontief, "Quantitative Input and Output Relations in the Economic System of the United States," *Rev. Econ. Statist.*, Aug. 1936, *18*, 105-25.

——, *The Structure of the American Economy, 1919-1939*, Oxford 1941.

A. P. Lerner, "Statics and Dynamics in Socialist Economics," *Econ. J.*, June 1937, *47*, 253 70.

——, "Theory and Practice in Socialist Economics," *Rev. Econ. Stud.*, Oct. 1938, *6*, 71-75.

E. Malinvaud, "Croissances optimales dans un modele macroeconomique," in *The Econometric Approach to Development Planning*, Amsterdam 1965.

——, "Decentralized Procedures for Planning," in Edmond Malinvaud and Michael Bacharach, eds., *Activity Analysis in the Theory of Growth and Planning*, London, New York 1967.

Alan S. Manne, "Sufficient Conditions for Optimality in an Infinite Horizon Development Plan," *Econometrica*, Jan. 1970, *38*, 18-38.

—— **et al.,** *Investments for Capacity Expansion: Size, Location, and Time-Phasing*, Cambridge, Mass. 1967.

—— **and Harry Markowitz,** *Studies in Process Analysis*, New York 1963.

P. Massé and R. Gibrat, "Application of Linear Programming to Investments in the Electric Power Industry," *Manage. Sci.*, Jan. 1957, *3*, 149-66.

James E. Meade, *Trade and Welfare*, Oxford 1955.

V. V. Novozhilov, "Metody soizmerenia narodnokhaziaistvennoi effektivnosti planovikh i proiektnikh variantov," ("Methods of Comparison of the National Economic Efficiency of Plan- and Project-Variants"), *Transactions of the Leningrad Industrial Institute*, 4, 1939.

J. D. Pitchford, *Population in Economic Growth*, Amsterdam 1974.

——, "Population and Economic Growth: Macroeconomics," paper presented at the Econometric Society, Third World Congress, Toronto 1975.

F. P. Ramsey, "A Mathematical Theory of Saving," *Econ. J.*, Dec. 1928, *38*, 543-59.

P. A. Samuelson, "Market Mechanisms and Maximization," Rand Corp., 1949; reprinted in *The Collected Scientific Papers of Paul A. Samuelson*, Cambridge, Mass. 1966, 425-93.

Adam Smith, *The Wealth of Nations*, 1776.

A. W. Tucker, "Linear and Nonlinear Programming," *Oper. Res.*, Apr. 1957, *5*, 244-57.

J. von Neumann, "A Model of General Equilibrium," (trans. from German original of 1936, G. Morgenstern), *Rev. Econ. Stud.*, No. 1, 1945, *13*, 1-9.

L. Walras, *Eléments d'économie politique pure*, Lausanne 1874; trans. W. Jaffé, *Elements of Pure Economics*, London 1954.

Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South

By ROBERT W. FOGEL AND STANLEY L. ENGERMAN*

In 1968 we undertook to measure and explain the relative technical efficiency of input utilization in the agricultural sectors of the North and South in 1860. The principal instrument that we employed for this task was the geometric index of the relative total factor productivity, which is defined by equation (1) (symbols are defined in Table 1):

$$(1)\quad G_s/G_n = \frac{Q_s/Q_n}{(L_s/L_n)^{\alpha_L}(K_s/K_n)^{\alpha_K}(T_s/T_n)^{\alpha_T}}$$

This index[1] was originally computed from published census data and the results were reported in 1971, both with and without adjustments for differences in the quality of outputs and inputs. The ratio of G_s/G_n yielded by the unadjusted computation was 109.2.[2] Crude adjustments for differences between the weights of northern and southern livestock, for land quality, for the proportion of women and children in the labor force, and for other factors, did not reduce this ratio as we thought they would, but increased it to 138.9.[3]

All differences between the northern and southern indexes of total factor productivity are, in a certain sense, errors of measurement. If output was correctly measured, and if all the inputs and conditions of production were fully specified and correctly measured, the ratio G_s/G_n would be equal to 100. To explain why G_s/G_n deviates from 100, then, is a process of accounting for such errors of measurement as omitted inputs, failure to adjust for differences in the quality of inputs, neglect of economies of scale or of improvements in the organization of production, omitted outputs, disequilibria in markets, and differences in product mixes.[4]

*Professor of economics and history, Harvard University, professor of American history and institutions, University of Cambridge; professor of economics and history, University of Rochester, respectively. Research on this paper was supported by National Science Foundation grants GS-3262, GS-27262, SOC 76-002, and a grant from Harvard University. Earlier versions were presented to faculty colloquia and seminars at the London School of Economics and the universities of Cambridge, Aberdeen, Uppsala, Oslo, Glasgow, Texas A&M, Oxford, Cologne, Berlin (The Free University), Moscow, Edinburgh, Jerusalem, and Tel Aviv. We benefited from the discussions following these presentations as well as from comments and criticisms by D. G. Champernowne, Michael Edelstein, Ephim Fogel, Michael Fogel, Robert Gallman, Zvi Griliches, Mark Hopkins, Laurence Kotlikoff, David Landes, Donald McCloskey, W. B. Reddaway, Joseph Reid, Jr., Richard Rosett, Allan Sanderson, Richard Sylla, James Trussell, and G. Nicholas von Tunzelmann.

[1]In the computations that follow, this index is multiplied by 100 to put it in percentage form. Hereafter we refer to G_s/G_n as the geometric index of total factor productivity, or the productivity index, or the index of efficiency. See the authors (1974a) (hereinafter referred to as *TOTC*), II, pp. 126–31.

[2]See Table 1 for brief descriptions of the way in which Q, L, K, and T were measured in the various computations. For further details see the authors (1971) and *TOTC*, II, pp. 131–36.

[3]Because of differences in the values of the α_i, the values of G_s/G_n presented in Tables B.20 and B.21 of *TOTC*, II, were 106.4 and 140.8. In the 1971 article the factor shares were $\alpha_L = 0.60$, $\alpha_K = 0.20$, and $\alpha_T = 0.20$. In *TOTC* the corresponding factor shares, which were derived from a production function for the cotton South by dividing the output elasticities by $1 + \sigma$, were 0.58, 0.17, and 0.25. See fn. 8, below.

[4]To restate the point more formally, equation (1) assumes that the production function in the ith region is $Q_i = f(L_i, K_i, T_i)$ rather than $Q_i = f(L_iA_{iL}, K_iA_{iK}, T_iA_{iT})$, where A_{iL}, A_{iK}, A_{iT} are the factor-augmenting coefficients in the ith region needed to transform the inputs of labor, capital, and land into "efficiency units." Note that in the augmenting formulation, differences in rainfall or sunshine between regions do not necessarily imply different regional production functions. As long as the output elasticities are the same for both regions, and there are no problems in aggregating output, the production functions will be the same even though the factor-augmenting coefficients differ. However, to the extent that the greater rainfall and sunshine of the South raises A_{sT} relative to A_{nT}, the failure to convert the land input in

In order to measure the effect of slavery on the process of production, it is therefore necessary to distinguish those mismeasurements that represent specific features of the slave system from those that are due merely to imperfections in the data, imperfections in methods of aggregation, or other mismeasurements that have no particular bearing on the operation of the slave system. In other words, we wish to obtain a residual measure of efficiency limited exclusively to measurement errors called "specific features of slavery." We then have the further task of identifying which specific features of slavery account for what parts of the aggregate value of the residual.

In our 1971 paper we stressed that a higher productivity index for the South than for the North did not necessarily imply that the southern advantage was due to special features of the slave system. We thought it was possible that slave-using plantations were less efficient than those using free labor, but that for some still undisclosed reason free southern farms were extraordinarily efficient. The high value of the southern productivity index would then be the consequence of averaging over a high index for free farms and a low index for slave plantations. Another possibility was that both slave and free farms that engaged in diversified agriculture were about as efficient as free farms in the North but plantations specializing in the export staples were highly efficient. In that case the relative productivity of the South might be due not to slavery per se, but merely to an unusually favorable market situation in 1860 for those export staples that happened to be produced by slave labor.

While we did not at that time rule out these alternatives, evidence in the 1860 Census indicated that the large slave plantations produced not only more cotton per capita but also more food per capita than small free farms in the South. It therefore seemed likely that the relative efficiency of southern agriculture was probably related to certain special features of the slave system. We conjectured that two features of slavery were particularly important. The first is that labor, and perhaps other inputs, were employed more intensively under the system of slavery than under the system of farming with free labor. There is much testimony for the proposition that slaves worked more days per year and, perhaps, more hours per day than free farmers. Since our efficiency indexes measured the labor input not in man-hours but in man-years, the more intensive utilization of labor shows up not as greater labor input but as a higher level of productivity. In 1971 we were inclined to believe that our failure to take account of the greater number of hours worked per year by slaves than by free men explained all, or nearly all, of our index of the superior efficiency of slavery. We also considered the much-debated possibility that there were economies of scale in the slave sector of agriculture. Even scholars who thought that slave labor was less efficient than free labor had suggested that the lower quality of labor might have been offset by the superior entrepreneurship associated with large-scale plantations.

To test these hypotheses we launched a search for additional data. A sample of 5,700 estates containing information on the price, age, sex, skills, and handicaps of slaves was retrieved from the probate records of southern courts. Southern archives yielded a sample of the business records of roughly 100 large plantations containing

---

both regions into units of equal efficiency, as is the case when equation (1) is employed, raises G_s relative to G_n even when the production functions are identical

When the output elasticities, and hence the factor shares, in the North and the South differ, an index number problem arises, since the same factor shares must be applied to the North and the South in order to prevent the unit of measurement from affecting the result. If the factor shares of the two regions do not differ greatly, the index number problem will be minor. As we pointed out in our 1971 article, the value of G_s/G_n is robust to plausible alternative estimates of the factor shares. Had we reduced the labor share from 0.58 to 0.5, and raised the capital and land shares proportionately so that they summed to 0.5, the partially adjusted index of G_s/G_n would have risen from 138.9 to 147.6. Had we made the labor share 0.7, while reducing the capital and land shares proportionately, the value of G_s/G_n would have declined from 138.9 to 132.2. The issues posed by differences between the northern and southern product mixes are discussed in Section IV, below.

TABLE 1—DEFINITIONS OF SYMBOLS USED IN EQUATIONS AND TABLES

Q = output. For unadjusted G_s and G_n, both Q_s and Q_n were based on the quantities reported in the 1860 Census of agriculture. The factors for seed and feed as well as the ratios for converting the stock of livestock into annual production were obtained from Towne and Rasmussen. Uniform national prices of 1860, again taken from Towne and Rasmussen, were used to aggregate the quantities produced in each region into Q_s and Q_n. In partially adjusted and more fully adjusted G_s, the output of animal products was reduced to take account of the low slaughter weights prevailing in the South relative to northern slaughter weights.

L = input of labor. *Free Labor in the North* In unadjusted G_n, L_n was taken to be equal to the census count of farmers and other agricultural occupations listed in the 1860 Census plus 17 percent of males aged 10–15. Thus children under 10 and females were excluded from the labor force. Those included were counted as equivalent full hands. In partially adjusted G_n, females and children under 10 continued to be excluded and males 60 or over were also excluded. Males between 10 and 59 were converted into equivalent full hands, using the weights applied to male slaves in partially adjusted G_s. *Free Labor in the South* In unadjusted and partially adjusted G_s, the procedures were those employed in the corresponding northern indexes. In more fully adjusted G_s, free males age 15 or over were given the age-specific weights derived from the age-earning profiles of male slaves. On farms with 0–5 slaves, free females age 10 or over were given one half the corresponding age-specific weight for female slaves, and free boys aged 10–14 were given the same age-specific weight as male slaves of that age category. The weights applied to free females and children on plantations with over 5 slaves were reduced linearly as plantation size increased, reaching zero for plantations with 50 or more slaves. *Slaves* In unadjusted G_s, 83 percent of all slaves aged 10 or over were assumed to be in the labor force and all were counted as equivalent full hands. In partially adjusted G_s, a deduction was made for rural slaves employed as domestics rather than in agricultural production Rough estimates of age- and sex-specific weights based on reports of various authorities were used to convert males and females into equivalent full hands The weights employed for males were, ages 10 14, 0.40, ages 15–19, 0.88, ages 20–54, 1.0; ages 55 59, 0 75, ages 60 or over, 0. Weights for females at each age ranged between 70 and 78 percent of the corresponding weights for males In more fully adjusted G_s, the weights used to convert males and females into equivalent full hands were based on age-earnings profiles reported in (*TOTC*) *Time on the Cross* for slaves of each sex. No adjustment was made for females engaged in domestic service However, a deduction was made for males who were engaged in activities, primarily artisan crafts, not covered by the output measure

K = input of capital In unadjusted and partially adjusted G_s and G_n, K was measured as the annual rental value of livestock, implements and machinery, and buildings. This was taken to be equal to the value of these items multiplied by the rate of return on farm capital (10 percent) plus average annual rates of depreciation of 2 percent for buildings and 10 percent for implements and machinery In more fully adjusted G_s, K was measured by the value of implements and machinery, and a corresponding adjustment was made in G_n for purposes of comparison with more fully adjusted G_s

T = input for land. In unadjusted G_s and G_n, T was measured by total acres in farms. In partially adjusted G_s and G_n, T was measured by the value of land plus improvements. In more fully adjusted G_s, T was measured by the value of land plus improvements and buildings, and a corresponding adjustment was made in G_n for purposes of comparison with more fully adjusted G_s

$\alpha_L, \alpha_K, \alpha_T$ = shares in value of output of labor, capital, and land

α_i = output elasticities of the inputs
A = the intercept of the production function
G = geometric index of total factor productivity
σ = the scale factor $(\alpha_1 + \alpha_2 + \alpha_3 - 1)$
Y = the age of land (years of land settlement)
V = the value of land plus improvements
δ = the rate of land depletion
i = the rate of interest

I = number of improved acres
U = number of unimproved acres
γ_1 = price per acre of I
γ_2 = price per acre of U
s = a subscript denoting the South
n = a subscript denoting the North
$\hat{}$ = a "hat" over a variable denotes the logarithm of that variable

TABLE 2—INDEXES OF TOTAL FACTOR PRODUCTIVITY ON SOUTHERN FARMS, BY SUBREGION AND SIZE OF FARM (G_n = 100)

Size of Farm as Measured by the Number of Slaves Per Farm	Slave Exporting States (Old South)	Slave Importing States (New South)	All States in Parker-Gallman Sample (Cotton South)
0	98.4	112.7	109.3
1-15	103.3	127.2	117.7
16-50	124.9	176.1	158.2
51 or more	135.1	154.7	145.9
All slave farms	118.9	153.1	140.4
All farms (slave and nonslave) in the subregion	116.2	144.7	134.7

Source: *TOTC*, II, p. 139. The slave-exporting (Old South) states are Georgia, North Carolina, South Carolina, and Virginia; the slave-importing (New South) states are Alabama, Arkansas, Florida, Louisiana, Mississippi, Tennessee, and Texas.

either detailed information on the organization of production, including the daily activities of each slave in the labor force, or demographic information needed to adjust the labor input of women.[5] The data in these sources, combined with the data in the Parker-Gallman sample of over 5,000 southern farms listed in the manuscript schedules of the 1860 Census,[6] made it possible to refine the input and output measures of G_s. The net effect of these refinements was to reduce G_s/G_n to 134.7.[7] The new data also permitted the computation of total factor productivity indexes by farm size and subregion. Tables 2 and 3 show that the superior efficiency of southern agriculture was not due primarily to the high performance of the free farms of the South. Free farms of the Old South fell below the efficiency of northern farms by 2 percent, while free farms in the New South exceeded the efficiency of northern farms by 13 percent. Thus only 4 percent of the efficiency advantage of southern over northern agriculture was due to the superior performance of the free sector. Slave farms accounted for 96 percent of the southern advantage.

Table 3 shows that within each region

[5]A further description of these two samples, as well as of 21 additional samples containing evidence relevant to the analysis of the slave economy, is contained in the authors (1975), Table 2, pp. 6-8. The appendix of the same source (pp. 137-39) contains a state and county distribution of the 77,000 slaves in the probate sample, by plantation size. It also contains more extended descriptions of 6 of the other 22 samples.

[6]James Foust gives a detailed description of the design of the Parker-Gallman sample and of the information contained in it. He also tests various statistics computed from the sample against the aggregate 1860 Census to assess the representativeness of the sample.

[7]The evidence collected between 1971 and 1973 indicated that our previous belief that slaves worked more hours per year than free farmers was incorrect. We continued, therefore, to measure the labor input of slaves and free men in man-years, presenting only a brief and, in retrospect, inadequate justification for this procedure (see *TOTC*, I, pp. 207-08). The question is discussed more fully in Section III of this paper.

Paul David and Peter Temin (p. 277) have argued that the interim weights that we applied to labor in the North biased the productivity comparison in favor of the South rather than in favor of the North. However, the ratio of equivalent peak-productive hands to persons on farms in the North yielded by our partially adjusted procedure is 0.273. If we had applied the same age- and sex-specific weights in the North that we did for the free farmers in the South under the more fully adjusted procedure, the ratio of equivalent peak-productive hands to persons on farms would have been 0.389. In other words, if we had used the procedure proposed by David and Temin, the labor input of northern farms would have increased by 42.5 percent. As a consequence the ratio G_s/G_n would have been not 134.7, but 165.4.

TABLE 3—THE RELATIONSHIP BETWEEN TOTAL FACTOR PRODUCTIVITY AND FARM SIZE IN EACH REGION
(Index of Free Farms in Each Region = 100)

Number of Slaves Per Farm	Slave Exporting States (Old South)	Slave Importing States (New South)	All States in Parker-Gallman Sample (Cotton South)
0	100.0	100.0	100.0
1-15	105.0	112.9	107.7
16-50	126.9	156.3	144.7
51 or more	137.3	137.3	133.5
All slave farms	120.8	135.8	128.5

Source: *TOTC*, II, p. 139.

fficiency increased with farm size, except hat in the New South the efficiency index is igher for medium than for large plantaons.[8] While we considered the possibility ıat in the West this intermediate category f slave plantations was actually more effiient than large plantations, we believed hat the reversal was probably due to meaurement errors. One was a failure to adjust dequately for the locational component of ınd values, which might have accounted or a much larger share of total land value n slave plantations with 51 or more slaves, specially in the New South, than on slave lantations in the 16 50 category. Another as the inadequacy of our adjustment for mitted products. Large slave farms, espeially in the West, probably engaged much ıore heavily in home manufacture than did mall ones. Large slave farms also appear to have devoted a larger share of the labor force to domestic services than did small plantations. We did not think that when these adjustments were made the entire differential in efficiency between the Old and New South would disappear. The continuous flow of labor from the Old South to the New South suggests that the long-run equilibrium between the two regions had not been attained by 1860. Hence one would expect to find some efficiency advantage in the newer area.

The criticism of our efficiency computation have largely followed the lines of analysis set forth in both our 1971 essay and in *Time on the Cross*. The important and widely cited critique of our productivity measures by Paul David and Peter Temin, for example, is largely a restatement of the caveats that we incorporated in the presentation of our findings, except that in virtually every place where we said that the measurement bias was negligible or went against the South, David and Temin sought to make the case that it was large or went in favor of the South. In this effort, the critics did not present new bodies of evidence that we overlooked but either conjectured upon the possibility that missing evidence could be found that would overturn our findings or else dwelt upon what they considered to be internal inconsistencies in our analysis or between various parts of the evidential corpus. The debate over efficiency has thus focused on a set of issues

[8]Table 3 suggests economies of scale. To test for this ossibility we fitted

$$\hat{Q} = \hat{A} + \alpha_2(\hat{K} - \hat{L}) + \alpha_3(\hat{T} - \hat{L}) + (1 + \sigma)\hat{L}$$

› measures of the inputs and outputs derived from the arker-Gallman sample. The resulting regression was

$$\hat{Q} = 2.898 + 0.1815(\hat{K} - \hat{L}) + 0.2606(\hat{T} - \hat{L}) + 1.0645\hat{L}$$
$$\quad\quad (0.0113) \quad\quad (0.0125) \quad\quad (0.0124)$$

igures in parentheses are standard errors). The scale ıctor ($\sigma = 0.0645$) is significant at well beyond the 001 level. Similar regressions, fitted to data for varius regions, indicate that the scale factor was larger ı the New South than in the Old South. A more comlete discussion of these regional regressions will be resented in the authors' forthcoming book.

that both sides view as the agenda for a new round of research.[9] The balance of this paper reports on our progress in working through that agenda

I. The Use of 1860 Price Relatives and Cotton Output

In chapter 3 of *Time on the Cross* we presented evidence indicating that the 1850's were a boom period for cotton planters, with the demand for cotton increasing rapidly and cotton prices generally well above their long-run trend. These findings led Thomas Haskell, David and Temin, Gavin Wright, Harold Woodman, and a number of others to argue that the high relative efficiency of southern agriculture indicated by the ratio G_s/G_n is merely an artifact of the temporarily inflated price of cotton in 1860 of our decision to use 1860 price relatives instead of those of a more normal year in computing the aggregate output of both the North and the South.

Even if this argument about the direction of the bias were correct, there would still be the question of the magnitude of the bias Between 1857 and 1860 there was a very substantial supply response on the part of cotton producers, which led to a fall in prices from the 1857 peak. By 1860, cotton prices had declined to within 8 percent of their long-run trend, or equilibrium, value. If, in aggregating southern output, we reduced the price applied to cotton by 7.5 percent $[1 - (1/1.08) \approx 0.075]$, the ratio G_s/G_n would decline from 134 7 to 131.8.[10]

What has been overlooked by Haskell and others is that 1860 was a boom year for all of agriculture and not merely cotton Consequently, the direction of the bias introduced by using 1860 prices to aggregate output does not turn on whether cotton was high relative to its price in other years, but whether it was higher than normal relative

TABLE 4—THE RATIO OF THE PRICES OF THE PRINCIPAL PRODUCTS OF NORTHERN FARMS TO THE PRICE OF COTTON

	1860 (1)	1850 (2)	(3)[a]
Corn/cotton	4.0	3.4	118
Wheat/cotton	8.9	6.8	131
Hogs/cotton	42.5	27.5	155
Cattle/cotton	33.4	24.3	137

Source: Computed from data in Marvin Towne and Wayne Rasmussen, pp. 283, 284, 294, 297, 308. Cotton is in dollars per pound, hogs and cattle in dollars per hundredweight, corn and wheat in dollars per bushel.
[a]Columns (1) ÷ (2) × 100.

to other agricultural prices, particularly to those of the principal northern agricultural products. Table 4 shows that the prices of the principal northern products relative to cotton were *higher* in 1860 than in 1850 (the only other year in which the data are sufficiently complete to permit the type of inter- and intraregional efficiency comparisons that we require). Had we chosen the 1850 set of price relatives or computed our efficiency indexes for 1850, as some have advocated, we would have increased (not reduced) the measured efficiency of slave agriculture relative to that of free agriculture as well as the advantage of large plantations relative to small farms.

Is it possible that both 1850 and 1860 are rare exceptions and that the price relatives of most other years would in fact support the case of the critics? Table 5 provides an answer to that question. It shows that the ratio of all farm product prices to that of cotton was 18 percent higher in 1860 than it was on average over the entire half century from 1811 to 1860. Thus in choosing the price relatives of 1860 to construct the output indexes, we chose a set of prices less favorable to the South than those that prevailed in 34 of the 49 years preceding 1860.

Wright (1975) argued that the choice of 1860 biased the productivity computation in favor of the South for still another reason. He believes that the South was the beneficiary of "random fluctuations in yields around their normal levels" (pp. 449–50)

[9]Other important critiques of our efficiency measures include Lance Davis, Thomas Haskell (1974, 1975), Harold Woodman, and Gavin Wright (1974, 1975).

[10]The cotton share of southern farm output (0.284) was computed from data underlying the computations described in *TOTC*, II, pp. 131–38.

TABLE 5—THE RATIO OF AN INDEX OF ALL FARM PRODUCT PRICES TO THE PRICE OF COTTON
(ratio for 1811-20 = 100)

1860	135
Average for the Half Century	
1811-60	114
Decade Averages	
1811-20	100
1821-30	96
1831-40	115
1841-50	136
1851-60	143

Source Computed from U.S. Bureau of the Census, 1960, pp. 115, 124 For each time period the average value of the index of all farm product prices was divided by the average price of cotton. The resulting ratios were then expressed as a percentage of the ratio for 1811-20.

that made the cotton crop of 1860 unusually large. This fortunate event, he conjectured, caused cotton production in the New South to be between 35.2 and 43.5 percent above its predicted, or normal, level (p. 444). Wright presented no actual evidence on yields to support this conjecture. Since the U.S. Department of Agriculture (*USDA*) did not begin collecting data on cotton acreage until after the Civil War, there is no basis for systematic estimates of annual cotton yields per acre harvested during the antebellum years.

Over the years 1867–1900, for which data on cotton yields are available, the average annual yield per acre is 177.1 pounds and the standard deviation is 20.10.[11] Thus a chance deviation in yields of between 35 and 44 percent above the mean (between 3.1 and 3.8 standard deviations) is an event so rare (less than one in a thousand) that one would expect it to have occasioned great comment among planters and in the agricultural publications of the time. Yet while much was said about the large size of the 1860 crop, the available commentaries are devoid of references to an extraordinarily high yield (see, for example, U.S. Bureau of the Census, 1862, p. 84; Lewis Cecil Gray, chs. 30 and 37; James Watkins, p. 10).

Wright based his conjecture on a supply curve for cotton that he has estimated, in which the output of cotton in a given year is made a function of the price of cotton (lagged one year) and the cumulative sales of public land (lagged two years) in four states of the New South and Florida (Wright 1971, pp. 111, 114). When Wright argues that the 1860 production of cotton in these states was above the predicted level, he means that observed output was above the level predicted by a model in which only cumulated past sales of public land can shift the supply curve of cotton. He arbitrarily assigns the entire residual for 1860 to cotton yields, although at least some part, if not all of it, is merely the artifact of an equation that failed to take account of such other determinants of supply as the proportion of privately held land that was in farms, the proportion of land in farms that was improved, the proportion of improved land sown in cotton, and the labor to land ratio. Rather than being fixed during the decade of the 1850's, as Wright assumed, these ratios were changing in such a manner as to increase the supply of cotton more rapidly than the cumulative total of public land sales. Between 1850 and 1860 the amount of improved land in the farms of the four states of the New South singled out by Wright increased by 59 percent, which was 1.6 times more rapid than corn production, 8.4 times more rapid than sweet potato production, and 20 times more rapid than the stock of hogs (U.S. Bureau of the Census 1862, pp. 196–236). This suggests that cotton was getting an increasing share of improved land at the expense of the principal food and feed crops.[12] In several cases, in-

[11]Computed from *USDA*, 1955, p. 5. The observation for 1866 was dropped because yields in that year were abnormally low as a result of the postwar disorganization of southern agriculture. But its inclusion would not affect the argument.

[12]Regressions relating yield per acre of cotton and yield per acre of corn in the cotton states estimated for the period 1867–1900 indicate that the elasticity of corn yields with respect to cotton yields was 1.1. It follows that if weather caused the cotton yield of these states to be 35 percent above normal, one would expect corn yields to be 38 percent above their normal level. Since Wright argues that yields in 1850 were below

cluding oats and rice, output not only failed to keep pace with rate of growth of improved land but declined absolutely. Public land sales are, of course, irrelevant to the supply of cotton in states that accounted for over a quarter of the crop of 1860, since they were not public-land states. In South Carolina, for example, cotton production increased by 17 percent between 1850 and 1860, although land in farms decreased slightly (see U.S. Bureau of the Census 1862, pp. 196-209).

What then explains the big increase in the output of cotton between 1850 and 1860? A complete answer to this question is beyond reach, but factors explaining about 91 percent of the increase can be identified. While we do not know what *annual* cotton yields were prior to the Civil War, in 1868 the U.S. commissioner of agriculture asked his corps of crop reporters to ascertain the normal cotton yields, as well as the normal share of tilled land sown in cotton, that had prevailed in 1860 and the years immediately preceding it. The figures obtained by this survey (pp. 414-15) indicate that the shift in the geographic locus of cotton production from the Old to the New South explains about 8 percent of the increase in output between 1850 and 1860. Assuming that within each state cotton just maintained its share of improved land, the increase in improved land explains 41 percent of the growth in the cotton crop. The amount of land switched within states from other crops to cotton cannot be known with certainty. But if we assume that the entire shift was confined to corn, the failure of corn production to keep pace with the increase in improved land implies that within-state reallocations of improved acreage explain another 42 percent of the growth in the cotton crop. These estimates leave a residual of 9 percent to be explained by all other factors including increases in the use of fertilizers, increases in the labor to land ratio, and random fluctuations in yields.[13]

But even if one grants so fortunate an event in 1859-60 as cotton yields that were 1.65 standard deviations above the mean (an event expected only once in 20 years), and assuming that northern farmers shared none of the benefits of the favorable weather that supposedly visited the South during 1859-60, the appropriate adjustment would just reduce G_s/G_n from 134.7 to 128.7. If we grant both fortuitously high cotton yields and inappropriately high price relatives, the ratio G_s/G_n declines to 126.3.[14]

II. Problems in Measuring the Land Input

We turn now to the two principal criticisms of our measure of the land input. In

normal, if the acreage devoted to corn had been held constant, one would expect the random fluctuation of yields hypothesized by Wright to have made the 1860 output of corn at least 38 percent greater than in 1850. That corn output actually increased by only 19 percent, therefore leads to the improbable conclusion that there was not only a proportionate decline but an absolute decline of 16 percent in the acreage devoted to corn between the two dates.

[13]The increase in output due to interstate shifts has two components: that due to the increase in the Southwide proportion of land in cotton and that due to the increase in the Southwide average yield per acre. Data for estimating both components were obtained from U.S. Bureau of the Census, 1895, pp. 92, 100; and *USDA*, 1868, pp. 414-15. The reallocation of corn land to cotton within states was computed by subtracting the actual 1860 production of corn in each of the 10 cotton states from the crop that would have been observed in each state if the growth of output had kept pace with the increase of improved land. The sum of these differences divided by 13.350 bushels, the average corn yield per harvested acre in these states between 1867 and 1900 (*USDA*, 1954), gives the number of acres switched from corn to cotton (there was no trend in corn yields between 1867 and 1900). The 1860 output of cotton divided by 177.1 pounds, the average per acre yield of cotton between 1867 and 1900, gives the number of acres in cotton in 1860. The last figure divided into the acreage that would have been in cotton if corn land had not been switched to cotton, represents the percentage increase in the output of cotton due to the within-state shift of land into cotton. This computation assumes that none of the increase in corn production between 1850 and 1860 is due to random fluctuations in yields. As fn. 12 indicates, such a fluctuation implies a greater reallocation of corn land to cotton than we have allowed. The distribution of interaction terms was resolved by converting the calculation to annual rates of change.

[14]The reduction in price, of course, only applies to the output of cotton reduced by 15.8 percent, the reduction which corresponds to 1.65 standard deviations. To grant a random increase in yields of 1.65 standard deviations above the mean is to attribute more than 175 percent of the unexplained increase in the cotton crop to good fortune.

our quality adjustment of the land input we used land values (V) rather than the rental value of land [$(i + \delta)V$], which contains a term for land depletion. It has been argued that land was depleted in the South, especially in the slave-selling states, much more rapidly than in the North.[15] To test this hypothesis we examined the relationship of land yields to the length of settlement in the selling states.[16] The effect of the length of settlement on land yields was estimated from equation (2):

$$(2) \quad \hat{Q} - \hat{T} = \beta_0 + \beta_1(\hat{L} - \hat{T}) + \beta_2(\hat{K} - \hat{T}) + \beta_3 Y$$

The equation was fitted to output and input measures constructed from the Parker-Gallman sample. When T was measured by total acres, the resulting regression was

$$(3) \quad \hat{Q} - \hat{T} = 3.988 + \underset{(0.0243)}{0.6070}(\hat{L} - \hat{T}) + \underset{(0.0174)}{0.2682}(\hat{K} - \hat{T}) - \underset{(0.00077)}{0.00570}Y$$

$$R^2 = 0.5045; \quad N = 1{,}539$$

When T was measured by improved acres, the resulting regression was

$$(4) \quad \hat{Q} - \hat{T} = 3.336 + \underset{(0.0242)}{0.3736}(\hat{L} - \hat{T}) + \underset{(0.0167)}{0.1786}(\hat{K} - \hat{T}) - \underset{(0.00071)}{0.00558}Y$$

$$R^2 = 0.2589; \quad N = 1{,}539$$

In both regressions, β_3, which we interpret as the rate at which land yields changed with the length of settlement, is statistically significant, negative, but small. Whether this rate of decline, barely 1/2 of 1 percent per annum, was more or less than in the North remains to be determined. However, even if one assumes that the northern depletion rate was zero, an adjustment for depletion in the South would raise that region's input of land by less than 6 percent. This rise, in turn, would reduce G_s/G_n from 134.7 to 132.9. Combining the adjustment for depletion with those for high cotton prices and fortunate yields, reduces G_s/G_n to 124.5.

David and Temin centered their criticism of our adjustment for differences in land quality on our failure to remove the locational component of land values. This omission created a more serious problem than the neglect of depletion, especially for the intra-South efficiency comparisons. The problem would not arise if farm-gate prices had been used in constructing the output index. For then farms located far from markets would receive lower prices for their products and also have lower values per acre of land than farms close to markets. Since uniform national prices were employed in constructing the output index, we implicitly assumed that all farms used the same average amount of transport service. If, as is frequently asserted, large slave plantations were generally better located than small farms, the assumption would introduce a bias against large plantations in the productivity comparisons. To test this hypothesis we estimated equation (5), on farms of various sizes, both by state and for the South as a whole.[17]

$$(5) \qquad V = \gamma_1 I + \gamma_2 U$$

The preliminary results of the analysis are summarized in Table 6, which shows that the locational component of land (estimated as γ_2 per acre) for the South as a whole was an average of 44 percent of land values. Thus 56 percent of so-called land values were due to investment in the land, and not, as is so often argued, to locational rents. Plantations with 51 or more slaves, however, were an exception. These large plantations, it turns out, were much better

[15]See Cairnes, pp. 52–53; Phillips, pp. 332, 336; Gray, pp. 447–448; Genovese, pp. 85–105.

[16]See Wright (1969), ch. 4, for an earlier attempt to test this hypothesis. Wright did not, however, construct an aggregate index of output nor adjust for differences in the quality of the inputs of labor.

[17]See Wright (1969), p. 95. Wright's regressions were fitted to the counties in the various soil categories that were constructed by Gray. Wright did not attempt to produce state or Southwide estimates of γ_1 and γ_2 by farm size but he did point out that it was γ_1 that explained most of the value of farmland and improvements.

TABLE 6—THE SHARE OF IMPROVEMENTS IN THE TOTAL VALUE OF LAND PLUS IMPROVEMENTS, BY FARM SIZE

Number of Slaves per Farm:	0	1-15	16-50	51+	All Farms in Cotton South
(1) Improved acres per farm (I)	46.85	121.25	349.76	930.24	141.94
(2) Unimproved acres per farm (U)	140.41	315.69	653.37	1,710.85	320.72
(3) Price per acre of I in \$ (γ_1)	13.138	20.243	23.294	30.949	22.30
(4) Price per acre of U in \$ (γ_2)	2.572	2.533	2.931	12.613	4.288
(5) $\gamma_1 - \gamma_2$ (\$)	10.566	17.710	20.363	18.336	18.012
(6) $\gamma_2(I + U)$ (\$)	481.63	1,106.77	2,940.17	33,312.07	1,983.89
(7) $(\gamma_1 - \gamma_2)I$ (\$)	495.02	2,147.34	7,122.16	17,056.88	2,556.62
(8) Total value of land and improvements per farm (Row 6 + Row 7 in \$)	976.65	3,254.11	10,062.33	50,368.95	4,540.51
(9) Improvements as a share of total value of land and improvements (Row 7/Row 8, percent)	50.69	65.99	70.78	33.86	56.31
(10) Location as a share of total value of land and improvements (100 − Row 9, percent)	49.31	34.01	29.22	66.14	43.69

Source: Computed from data in the Parker-Gallman sample, using the values of γ_1 and γ_2 estimated from the Southwide regressions run on equation (7) for each size class. Values of γ_1 and γ_2 were, in all cases, significant at the 0.001 level. The entries in rows 1-4 of the all-farms column are weighted averages of the corresponding entries for columns 1-4. The weights in rows 1 and 2 are the proportion of farms. The weights in rows 3 and 4 are the proportion of improved and unimproved acres, respectively.

located than small ones. The average locational rent per acre on large plantations was more than four times that on smaller-sized farms, whether slave or free.[18] Consequently, even though investment per improved acre on large plantations exceeded

[18]It might be argued that better (more fertile) land was brought into production first. In that case $\gamma_1 - \gamma_2$ would measure not only the average investment per improved acre but also the superior quality of the land on which the improvement had been made. Even if this were the case all of $\gamma_1 - \gamma_2$ belongs in the quality adjustment, since whether this amount is due entirely to improvements or reflects some superior virgin quality, it is "more" land in an economic sense. We can test the hypothesis that more fertile land was improved first by making use of Martin Primack's estimates of labor required to improve an acre (pp. 154-55, 233-43), and annual wage rates on southern farms (*USDA* 1868, p. 416). These indicate that the average value of improvements per acre was \$20.05. Since $\gamma_1 - \gamma_2$ = \$18.01, Primack's data suggest that all of $\gamma_1 - \gamma_2$ is investment. Probably ease of land clearing rather than natural fertility was the principal factor determining which land was improved first.

It might also be argued that the land belonging to plantations with 51 or more slaves was more fertile than the land of smaller farms. If so, some part of γ_2 for this class would represent not locational advantage but the superior fertility of land yet to be brought into production, and the subtraction of all of γ_2 from γ_1 would underestimate the quantity of quality-adjusted land plus improvements that must be included in the land input of this class. Alternative assumptions regarding the distribution of γ_2 between locational advantage and superior fertility led to variations in the index of total factor productivity for this class of farms ranging from 147.7 to 132.8 (see Table 7, col. 3). Two aspects of this result should be stressed. First, even if we assumed that the locational advantage of farms with 51 or more slaves was no greater than that of farms with 16-50 slaves, none of the arguments in the balance of this paper would be altered. Second, the principal effect of variations in the distribution of γ_2 is on the assessment of the comparative efficiency of intermediate and large plantations. Assuming that plantations in the 51-or-more class had greater locational advantage than those in the 16-50 class leads to the conclusion that they were more productive than those in the 16-50 class. If we assumed that locational advantages were identical, then both classes would have roughly the same indexes of total factor productivity.

TABLE 7—THE EFFECT OF CORRECTING FOR THE LOCATIONAL COMPONENT OF LAND VALUES ON THE RELATIVE VALUES OF THE INDEXES OF TOTAL FACTOR PRODUCTIVITY
(Index of Free Southern Farms = 100)

Number of Slaves per Farm	Index Before Correction, All States in Parker-Gallman Sample	Index After Correction, All States in Parker-Gallman Sample
0	100.0	100.0
1 15	107 7	100.8
16 50	144.7	133.1
51 or more	133.5	147.7

the cotton-South average, investment accounted for only 34 percent of land values.[19]

It follows that the previous failure to correct for locational rents biased the efficiency index for plantations with 51 or more slaves downward, while the indexes for plantations with 1 15 and 16 50 slaves were biased upward. Table 7 shows that after correction for locational rents, total factor productivity on slave farms increases continuously with farm size, but small slave farms have no productivity advantage over free farms.

III. The Length of the Work Year

One of the most important results of the work over the past two years is the finding by Jacob Metzer and John Olson that slaves worked fewer hours per year than free farmers. This directly contradicts the hypothesis of our 1971 paper that the measurement of the labor input in man-years rather than man-hours was the principal reason that G_s exceeded G_n. The belief that slaves worked more days per year than free men is not only widely held but was recently reasserted by David and Temin (pp. 768-71) who attributed the phenomenon to the fact that there were more frost-free days in the South than in the North. They put the average length of the work year of northern farmers at 2,163 hours. Noting that there are between 220 and 240 frost-free days in the South but only between 160 and 180 such days in the North, David and Temin argued that it was reasonable to assume that slaves worked 10 percent more days per year than free northern farmers. Moreover, since the days during the winter were both longer and warmer in the South than in the North, they also reasoned that slaves probably averaged more hours per day than northern farmers.

Systematic data bearing on the average number of days worked per year and per season during antebellum times are available only for the South. Such averages have been computed by Metzer and Olson from the daily work records that were kept by the owners or overseers of slave plantations.[20] Processing of most of these is still in progress, and the computations employed in this paper are based on a subsample of 7 cotton plantations that are displayed in Table 8.

This table shows that there was relatively

[19]For the North γ_2 is just $1 80 per acre. Thus the locational component per acre was lower in the North than in the South. The result is surprising only if considered in a partial rather than a general equilibrium context Between 1840 and 1860 over 15,000 miles of railroad track were built in the North, bringing millions of acres close to a rail route. This massive increase in supramarginal land brought about by railroad construction probably lowered rather than raised the average locational rent (see Fogel, 1964, p 223, n. 10) According to Haskell (1975) our assumption that "an acre of northern farmland was an average of 2.5 times better in quality than southern farmland" is "extraordinary" (p. 38) But an "average" acre of northern farmland was superior to an "average" southern acre in 1860, not primarily because of its natural qualities, as Haskell appears to believe, but because more capital had been invested in its improvement. Over 54 percent of northern farmland had been improved by 1860 but for the South the corresponding figure is only 30 percent. Consequently, even if land in both regions had been of equal quality in the natural state and if expenditures per improved acre had been the same in both regions, the average northern acre would have been 1.8 times better than the average southern acre. But northerners invested about 30 percent more on each of their improved acres than did southerners. Thus about 90 percent of the "superior" quality of northern land represented investment. (Computed from data in Primack, pp. 154-55, 233-43; U.S. Bureau of the Census, 1895, pp. 84-100; Clarence Danhof, p. 77; and *USDA*, 1868, p. 416)

[20]The periods covered by these records range from a single season to several years. The detail contained is uneven, but it is possible to determine the daily work records of fieldhands and sometimes also that of artisans, servants, and others engaged in nonagricultural labor. It is also possible to determine holidays and days lost due to rain or illness.

TABLE 8—DAYS WORKED PER SEASON FOR A SAMPLE OF SOUTHERN COTTON PLANTATIONS

	Spring	Summer	Fall	Winter	Total
Prudhomme-Bermuda (LA)	71.6	72.1	65.6	61.9	271.2
Flinn-Green River (MS)	70.3	72.6	71.9	67.0	281.8
Monette-Hope and Pleasant Hill (LA)	73.9	74.5	67.1	65.6	281.1
Le Blanc (LA)	75.7	68.1	67.2	70.7	281.7
Pre Aux Cleres (LA)	73.5	69.5	68.6	68.2	279.8
El Destino (FL)	72.6	68.9	70.2	67.7	279.4
Average of six short-staple plantations (equal weights)	72.9	71.0	68.4	66.8	279.1
Kollock Ossabow Island (GA)	77.6	73.5	70.9	70.7	292.7
Average of seven plantations (equal weights)	73.6	71.3	68.8	67.4	281.1

Source: Olson.

little variation in the number of days worked per season, although the number of work days during the spring planting and cultivating period was about 7 percent greater than during the peak harvest months of the fall. The average number of days worked per year within the sample ranged from 271 to 293. The average number of days worked per year in the sample of 6 short-staple plantations is 279. Adding the long-staple Kollock plantation raises the average to 281 days.

Thus the number of days in the work year of slaves appears to have fallen short of the potential by about 23 percent. This result is explained primarily by the almost total absence of Sunday work.[21] Occasionally, a few hands were used on Sundays for special tasks. But such incidents were rare. This nearly total absence of Sunday work is a unique feature of the large slave plantations, and it bears on the special nature of the slave-labor system.

Unfortunately, information on the length of the agricultural workday for the antebellum era is fragmentary, although a recurrent theme is that the day extended from sunrise to sunset. The earliest systematic studies of regional and seasonal variation in the length of the workday carried out by the *USDA* pertain to the first third of the twentieth century. Olson points out that these studies yield seasonal estimates of the length of the workday that are quite similar to those obtained by subtracting standard time allowances for meals during antebellum times from the interval between sunrise and sunset in each region and season. Combining the information on the number of workdays presented in Table 8 with these seasonal estimates of the average length of the workday, Olson produced Table 9, which presents the number of hours in the slave work year for the sample of 7 cotton plantations. The total hours worked per year within the sample ranged from a low of 2,709 to a high of 2,912. The average for the entire sample is 2,798 hours.[22] Since the last figure is 29 percent *greater* than the 2,163 hours that David and Temin hold was typical on northern farms, and 18 percent greater than their estimate of the length of the slave work year, this finding might seem to substantiate the proposition that the slave work year exceeded the free one.

But such a conclusion is warranted only if one accepts the contention that the average work year of antebellum farmers in the North was just 2,163 hours. David and Temin cite a *USDA* report by John Hopkins on changing conditions of agricultural

[21]The balance of the shortfall is explained by other holidays and half-days on Saturdays (6 days), by illness (11 days), and by rain and inclement weather (15 days).

[22]Ralph Anderson, using a different sample of plantations and working independently of Olson, produced an estimate of the length of the slave work year that is quite similar to Olson's.

TABLE 9—HOURS WORKED PER SEASON AND PER YEAR FOR A SAMPLE OF SLAVE COTTON PLANTATIONS

	Spring	Summer	Fall	Winter	Average per Season	Total
rudhomme-Bermuda	759	771	702	477	677	2,709
linn-Green River	745	777	769	516	702	2,807
Monette-Hope and Pleasant Hill	783	797	718	505	701	2,803
Le Blanc	802	729	719	544	698	2,794
Pre Aux Cleres	779	744	734	525	696	2,782
El Destino	770	737	751	521	695	2,779
Average of six short-staple farms	773	760	732	514	695	2,779
Kollock (long staple) Ossabow Island	823	786	759	544	728	2,912
Average of seven cotton farms	780	763	736	519	700	2,798

Source Olson

labor and technology between World War I and 1936 as the source for this figure. But Hopkins put the length of the northern work year during this period at between 2,800 and 3,370 hours (pp. 23, 27). Combining the samples reported by Hopkins with those of several other *USDA* studies, Olson computed the average length of the agricultural work year in the various subregions of the North. Although there was a fair degree of variation from sample to sample, as well as by the principal subregions, the average of every sample exceeded the 2,163 hours asserted by David and Temin by at least 31 percent. The lowest subregional average of 3,006 hours was found in the corn and general farming belt; the highest was 3,365 hours in the western dairy region. The average for the overall sample of 1,605 northern farms was 3,130 hours.

Comparison of these figures with Table 9 reveals that slaves worked approximately 10 percent fewer hours than northern farmers.[23] The contention that the number of frost-free days (or the length of the growing season) was the principal factor determining the length of the work year is, thus, incorrect. While the number of frost-free days determines which plants can be raised in a particular region, there is little relationship between the length of the growing season

[23]This conclusion, of course, rests on the assumption that the northern work year was not shorter in antebellum times than during the first third of the twentieth century. The assumption is consistent with available fragmentary data. See Olson and the sources cited there. The discussion of the length of the northern work year involves two issues. First, since David and Temin turned to *USDA* data on hours worked by farmers during the first third of the twentieth century, there is the question of the average length of the work year indicated by this body of evidence. The second issue is whether or not an average derived from data for the early twentieth century may be applied to the antebellum period, given the changes that occurred in income and in the technology of feeding, milking, planting, and harvesting. Gallman has pointed to factors which suggest that hours worked in northern agriculture may have been longer in the early twentieth century than in the mid-nineteenth. These are discussed in some detail by Olson. Clearly, a continued search for evidence in antebellum sources bearing on this question is called for But even if we granted the contention that the slave work year was 10 percent longer than the northern work year G_s/G_n would fall from 134.7 to 127.5 Combining this adjustment with the previous ones for high cotton prices, fortunate yields, and land depletion, reduces G_s/G_n to 117.9. So even if these conjectured corrections of southern inputs and outputs advanced by the critics were all correct, the efficiency of southern agriculture would still exceed that of northern agriculture by 18 percent. There is, of course, also the adjustment to the input of northern labor proposed by David and Temin. But as we pointed out in fn. 7, the direction of this adjustment is opposite to that conjectured by the critics. If we made it, G_s/G_n would rise from 117.9 to 144.8, thus wiping out the effect of the proposed adjustments to southern inputs and outputs.

and the duration of the period from seedtime to harvest for particular crops. The growing season in South Dakota, for example, is about 150 days but the period from seedtime to harvest is 310 days for winter wheat and only 115 days for spring wheat (see James Covert pp. 35, 36, 43, 44).[24]

The length of the work year was determined not only by the duration of the planting, cultivating, and harvest seasons of a particular mix of crops but also by the mix between field crops and animal products (including dairy products). Olson stresses that the length of the northern work year was positively correlated with the degree of specialization in livestock and dairying. The implications of the labor-intensive methods of rearing livestock (which was already characteristic of the North by 1860)[25] and of dairying are revealed by a *USDA* study showing that the average duration of the workday increased by over an hour on both weekdays and Sundays in counties that switched from general farming to dairying (see Olson; Hopkins, pp. 26–27). The principal reason for the longer work year in the North than on slave plantations is that the North specialized in dairy and livestock while slave plantations did not. Olson estimates that about 38 percent of the product of northern farms in 1860, as measured by value-added, originated in livestock and dairying. The corresponding figure for the cotton South is just 9 percent; for large plantations it is hardly 5 percent.

The finding that the slave work year was shorter than the free work year does not contradict the proposition that slave labor was more intensely exploited than free labor, but only the proposition that such exploitation took the form of more hours per year. What had been insufficiently emphasized is the possibility that slaves worked more intensively per hour than did free men.

[24]The relationship between the growing period, the growing season, and the period between seedtime and harvest are clarified by Covert, p. 14

[25]See Fogel (1965), p. 216. Corn consumption per equivalent hog more than doubled in the North between 1840 and 1860 but remained relatively constant in the South. See Percy Wells Bidwell and John Falconer (pp. 393, 437), Gray (pp. 842–45), and Gates (pp. 199, 218).

IV. The Problem of Differences in Product Mix

There is still the problem of the regional differences in the mix of products. Because the output mixes of northern and southern agriculture differed, the attempt to compare G_s and G_n poses an index number problem. This is an issue which, of course, plagues all productivity comparisons, whether over space or time. David and Temin argue that in our case the problem is insuperable because cotton could be grown only in the South, not in the North. Fortunately, the problem is not quite so intractable.

The influence of product mix can be tackled by taking advantage of the fact that there was a class of free farms that could, and did, produce cotton. While it is true that free southern farms were slightly more efficient than northern farms, this edge explained only 4 percent of the ratio $(G_s - G_n)/G_n$. Thus nearly all of the southern productivity advantage is explained by the extent to which the productivity of medium and large slave plantations exceeded the efficiency of small free farms, whether these small farms were located in the North (where they could not produce cotton) or in the South (where they did produce cotton).

To put the issue somewhat differently, while both the large slave and small free farms of the South produced cotton, the large slave plantations were 48 percent more efficient than small free farms (see Table 7). Of course, the cotton share of output on these small farms (29 percent) was less than that of large slave plantations (61 percent). But since there was no climatic obstacle that prevented the free southern farms of the cotton belt from choosing exactly the same mix of products that was selected by large slave plantations, it may be assumed that they chose the product mix that was most efficient for them. Presumably a product mix with a larger cotton share would have decreased, or at least not increased, their efficiency. In other words, it appears

that it was the manner in which cotton was produced, rather than the mere capacity to produce cotton, that is the principal factor accounting for the superior efficiency of large plantations. This interpretation is supported by the finding that small slave plantations, those with 1–15 slaves, were not more efficient than free southern farms (see Table 7) even though their cotton share (39 percent) was a third greater than that of free farms.

The last point needs elaboration, since Gavin Wright (1974, 1975) has argued that free farmers in the cotton belt were just as efficient in the production of cotton as large plantations but that they chose to produce less cotton in order to reduce riskiness. The basic propositions of his argument are that both small free and large slave farms were equally efficient in the production of cotton, that both were equally efficient in producing other commodities, and that the difference in mix between cotton and other commodities alone explains the differences in the indexes of total factor productivity by farm size. If Wright's conjectures are correct, the overall efficiency index of farms of each size class would be given by geometric averages of the indexes for cotton and other commodities, with the weights on each invariant commodity index being the share of that commodity in total output—the shares varying with size classes (see Evsey Domar).

If we let A_c equal the efficiency index for cotton and A_0 equal the efficiency index for all other commodities, the relationship of the overall efficiency index in each size class to A_c and A_0 is given by the following four equations:

(6a) $A_c^{0.29} A_0^{0.71} = 1.00$ (0 slaves)

(6b) $A_c^{0.39} A_0^{0.61} = 1.01$ (1–15 slaves)

(6c) $A_c^{0.53} A_0^{0.47} = 1.33$ (16–50 slaves)

(6d) $A_c^{0.61} A_0^{0.39} = 1.48$ (51+ slaves)

Since there are only two unknowns, only two equations are needed for a solution. If A_c and A_0 were invariant with size class, or approximately so, the ratio between A_c and A_0 should be approximately the same, regardless of which two equations are used to solve for A_c and A_0. However, as the following classification shows, this is not the case. The ratio A_c/A_0 varies from 1.1 to 7.1.

Equations used to solve for A_c and A_0	Ratio of A_c to A_0
(6a) and (6b)	1.1
(6a) and (6c)	3.3
(6a) and (6d)	3.4
(6c) and (6d)	3.8
(6b) and (6d)	5.7
(6b) and (6c)	7.1

In other words, it was not the mere difference in product mix but the manner in which cotton was produced that made large slave plantations more efficient than either small free farms or small slave farms. The threshold size for the efficiency of the gang system appears to be above 15 slaves.[26]

There is another aspect to Wright's argument. That is the contention that small free farms chose the observed mix in order to reduce riskiness—to reduce the variance of their income. In order to explore the implications of this suggestion, let us for the moment accept Wright's assumption that both large slave and small free farms were equally efficient in producing cotton. In that case, as we have seen, the income yield per average unit of input was more than three times as large in the production of cotton as in other products on farms of both sizes.[27] Consequently, by increasing the cotton share of their output from 29 to 61 percent (the share on large slave plantations), small farmers could have increased their mean income by 48 percent.

Wright's conjecture implies, therefore,

[26]This is not to say that the level of the overall efficiency index of each size class was independent of the share of cotton in total output. Quite the contrary, farms that were more efficient in cotton production should have been more heavily specialized in that crop. In other words, the optimum share of cotton in total output was a function not only of relative prices but of a farm's comparative advantage in that crop. Moreover, the mix of crops was a major determinant of the degree to which the labor force was utilized (see Section V, below).

[27]This result is obtained by solving equations (6a) and (6d) for A_c and A_0.

that farmers were willing to forgo close to half of their mean income in order to gain some unspecified reduction in the variance of that income. Not only is this an extraordinary price to pay for insurance, but there is no evidence to show that within the relevant range (cotton shares between 29 and 61 percent of gross farm product) the relative variance of income was positively correlated with the share of income originating in cotton. The absence of a positive correlation between cotton shares and the relative variance is indicated by the facts that yields in cotton were no more variable than those of other farm products and that the price of corn, which Wright argues was the principal substitute for cotton, was actually more variable than the price of cotton.[28] Nor do the benefits of a more mixed "portfolio" seem to have been potent enough to have provided less variation in price than that associated with cotton. Over the years 1831-60, Thomas Senior Berry's index (p. 564) of the average price of 20 or more agricultural products "identified" with the North (i.e., excluding cotton, sugar, and rice) has a coefficient of variation (0.23) that is quite similar to that for cotton alone (0.25).[29]

Wright has also argued that the objective of small farms may have been not the minimization of the variance of their income but "safety first." The exact meaning of safety first is never clearly defined, but Wright appears to equate it with the guaranteeing of the food supply. Thus before turning to the market with all its risks to obtain products that they did not produce, free farmers first sought to insure that they would not starve. However, as we showed in our analysis of the slave diet (1974b), free southern farmers could have guaranteed a nutritious diet, not only high in protein but exceeding all other nutrient requirements, out of their own production for less than six cents per capita per day. Thus it took less than one-third of the average annual product of a free farm to insure an adequate diet for all those living on such farms. If the free farms of the cotton belt were as efficient in cotton production as were the large plantations, they could have had both an insured food supply and a 48 percent increase in their mean income by raising the cotton share of their output from 29 to 61 percent.[30]

V. Sources of Efficiency on Large Plantations

The finding that large slave plantations were 48 percent more productive than the small free farms of the South poses a new problem: What feature, or features, of the organization and operation of large slave plantations gave them such a marked advantage? Examination of the managerial records of these plantations suggests that part of the answer lies in the persistence with which planters sought to exploit complementarities and interdependencies made possible by concentration in the production of one or the other of the four principal slave crops: cotton, sugar, rice, and tobacco.

[28]The relative variance of incomes on free and slave farms in 1860 was computed from the Parker-Gallman sample by weighting outputs on farms of all sizes by uniform national prices. The coefficient of variation was between 30 and 40 percent larger on free farms (with an average cotton share of 29 percent) than on slave farms (with average cotton shares ranging between 39 and 61 percent). Over the years from 1867 to 1900 the coefficient of variation in corn yields fell below that of cotton yields by just 0.02 (*USDA*, 1954, 1955). The relative variance of the price of corn in New Orleans over the years 1840-60 was 25 percent greater than the relative variance of cotton prices over the same period (computed from Cole).

[29]For the period 1831-46 there are 20 commodities in Berry's index for northern agriculture. For the period 1846-60, the number of commodities is 29. Cotton prices are from U.S. Bureau of the Census, 1960, p. 124.

[30]To farmers in debt, safety-first could well have meant guaranteeing a high enough cash income to meet mortgage and other debt calls. The point at issue is not whether the food supply mattered to antebellum farmers but whether they perceived it as the most binding constraint in the determination of their economic decisions. Illiquidity could well have appeared as a more serious menace than an inadequate food supply. Wright does not explore this issue, although it is widely suggested in the traditional literature on nineteen-century agriculture. Moreover, if A_c exceeded A_0 by 3.4 times, and if farmers sought to insure some minimum absolute level of income, then given the relative variances of income from cotton and corn, the much higher mean income associated with specialization in cotton indicates that the choice of the higher cotton share would have reduced rather than increased the likelihood of falling below that minimum.

The central focus of planters was the organization of the labor force into highly coordinated and precisely functioning gangs characterized by intensity of effort. "A plantation might be considered as a piece of machinery," said Bennet H. Barrow (Edwin Adams Davis, p. 409) in his Highland Plantation rules. "To operate successfully, all its parts should be uniform and exact, and its impelling force regular and steady." "Driving," the establishment of a rigid gang discipline, was considered the crux of a successful operation. Observers, such as Robert Russell, said that the discipline of plantation life was "almost as strict as that of our military system" (p. 180). Frederick Law Olmsted described one instance in which he observed two very large hoe gangs "moving across the field in parallel lines, with a considerable degree of precision." He reported that he repeatedly rode through their lines at a canter with other horsemen, "often coming upon them suddenly, without producing the smallest change or interruption in the dogged action of the labourers" (p. 452).

Each work gang was based on an internal division of labor that not only assigned every member of the gang to a precise task but simultaneously made his or her performance dependent on the actions of the others. On the McDuffie plantation, the planting gang was divided into three classes which were described in the following way (as quoted by Metzer):

> 1st, the best hands, embracing those of good judgment and quick motion. 2nd, those of the weakest and most inefficient class. 3rd, the second class of hoe hands. Thus classified, the first class will run ahead and open a small hole about seven to ten inches apart, into which the second class drop from four to five cotton seed, and the third class follow and cover with a rake. [p. 135]

Interdependence and tension were also promoted between gangs, especially during the period of cultivation when the field labor force was divided into plow gangs and hoe gangs. The hoe hands chopped out the weeds that surrounded the cotton plants as well as excessive sprouts. The plow gangs followed behind, stirring the soil near the rows of cotton plants and tossing it back around the plants. Thus the hoe and plow gangs each put the other under an assembly-line type of pressure. The hoeing had to be completed in time to permit the plow hands to carry out their tasks. At the same time the progress of the hoeing, which entailed lighter labor than plowing, set a pace for the plow gang. The drivers or overseers moved back and forth between the two gangs, exhorting and prodding each to keep up with the pace of the other, as well as inspecting the quality of the work. In operations such as cotton picking, which did not lend themselves as naturally to interdependence as planting and cultivating, planters sought to promote intensity of effort by dividing hands into competing gangs and offering bonuses on a daily and weekly basis to the gang that picked the most. They also made extensive use of the so-called "task" methods. These were, literally, time-motion studies on the basis of which a daily quota for each hand was established.

In addition to the use of assembly-line methods and time-motion studies to insure maximum intensity of effort in a particular operation, planters sought to allocate their slaves among jobs in such a manner as to achieve "full capacity" utilization of each person. In this connection slaves were given "hand" ratings—generally ranging from one-eighth to a full hand—according to their age, sex, and physical ability. The strongest hands were put into field work, with the ablest of these given tasks that would set the pace for the others. Plow gangs were composed primarily of men in their twenties or early thirties. Less sturdy men and boys, as well as prime-aged women, were in the hoe gangs. Older women were occupied in such domestic duties as house servants and nurses; older men worked as gardeners, servants, and stock-minders. Metzer's analysis (p. 134) of the records of the Kollock plantations indicates that the "hand"-to-slave ratio was 0.9 in field work but only 0.6 in nonfield work. Metzer points out that in allocating slaves among jobs, planters pursued the

TABLE 10—COTTON-PICKING RATES OF PREGNANT WOMEN AND NURSING MOTHERS AS A PERCENTAGE OF THE COTTON-PICKING RATES OF WOMEN THE SAME AGE WHO WERE NEITHER PREGNANT NOR NURSING

Weeks Before (−) or After (+) Childbirth	Age			
	20	25	30	35
−12 to −9	82 3	83.3	84.1	84.8
−8 to −5	77.4	78.8	79.8	80.6
−4 to −1	74.8	76.3	77 4	78.3
+2 to +3	3.9	9.8	14.1	17.4
+4 to +7	64 9	67.1	68.6	69.8
+8 to +11	91.3	91.8	92.2	92.5

principal of comparative, rather than absolute, advantage. Thus during the harvest period, most of the labor of picking cotton was provided by women, even though women had lower daily cotton-picking rates than men. On the Pleasant Hill plantation, for example, women provided 31 percent more labor time in cotton picking than did men.

Data on the cotton-picking rates of pregnant women and nursing mothers provide still another illustration of the degree to which planters succeeded in utilizing all those in the labor force. Estimates derived from Metzer's regression of the daily cotton-picking rates of women by age, and by weeks before and after childbirth, are summarized in Table 10. This table shows that down to the last week before birth, pregnant women picked three-quarters or more of the amount that was normal for women of corresponding ages who were neither pregnant nor nursing. Only during the month following childbirth was there a sharp reduction in the amount of cotton picked. Some mothers started to return to field work during the second or third week after birth. By the second month after birth, picking rates reached two-thirds of the level for nonnursing mothers. By the third month, the level rose to over 90 percent.

Another way in which planters sought to achieve full capacity utilization of labor was in the selection of the product mix. Labor requirements in cotton production had a very marked seasonal pattern, with one peak reached in the late spring and a second in October. Consequently, secondary crops were chosen so that their peak labor requirements were complementary to those of cotton (see Metzer, Figure 1). Corn was an excellent match. It could be planted before cotton and could be harvested either early or late, depending on other pressures, because the kernels, protected in the ears, did not suffer if harvesting was delayed beyond maturation.

VI. Some Issues of Interpretation

It should not be assumed that slave labor was more efficient than free labor in all occupations. There is no evidence that the productivity of slave labor exceeded that of free labor in urban industries. As Claudia Goldin (pp. 104-05) points out, the much higher elasticity of demand for slave labor in the cities than in rural areas (the ratio is more than 10 to 1) indicates that whatever advantage there was in slave labor was specific to agriculture.

Preliminary analysis also suggests that within *U.S.* agriculture, the slave system of labor raised productivity only for slave farms that specialized in one of four principal products: sugar, cotton, rice, and tobacco. Economies of scale seem to have been greatest in sugar, since nearly 100 percent of all cane sugar in the United States was produced on large slave plantations. The slave system seems to have been less productive in tobacco than in cotton. The scale factor in the Old South (where tobacco was a relatively important crop), while statistically significant, was only a third as large as the scale factor for the New South. Although the issue is now under investiga-

ion, there appears to have been no produc-ivity advantage to slave labor in general 'arming and relatively few large slave plan-ations engaged in general farming.

The available evidence indicates that ;reater intensity of labor per hour, not nore hours of labor per day nor more days)f labor per year, is the reason why the in-lex of total factor productivity is 48 percent iigher for slave plantations than for free 'arms. There is no evidence that land was ısed more intensively in the South than in he North, but even if it was, the depletion ate of southern land yields was so low (0.6)ercent per annum) that this can, at most, ıccount for 5 percent of the value of $G_s - G_n)/G_n$. The interim estimates thus ndicate that slaves employed on medium ınd large plantations worked about 72 per-:ent more intensively per hour than free 'armers. In other words, on average, a slave)n these plantations produced as much out-)ut in roughly 35 minutes as a free farmer lid in a full hour.

Once it is recognized that the fundamen-al form of the exploitation of slave labor vas through speed-up (increased intensity)er hour) rather than through an increase in he number of clock-time hours per year, :ertain paradoxes resolve themselves. The onger rest breaks during the work day, and he greater time off on Sundays, for slaves han for free men appear not as boons that ;lave-owners granted to their chattel but as :onditions for achieving the desired level of ntensity. The finding that slaves earned 15)ercent more income per clock-time hour is ess surprising when it is realized that their)ay per equal-efficiency hour was 33 percent ess than that of free farmers.

David and Temin argue (pp. 778-83) that vhile the index of total factor productivity nay be acceptable in comparing the relative :fficiency of free countries it cannot be ap-)lied to a comparison of the free North and he slave South—that in this instance a norally weighted index of efficiency is re-Juired. However, the issue of the relative :fficiency of slave labor did not originate vith *Time on the Cross*. It is an issue with a ong history that traces back to such com-nentators on slavery as Adam Smith, Alexis de Tocqueville, Cassius Marcellus Clay, Hinton Rowan Helper, Frederick Law Olmsted, and John Cairnes. Is the geo-metric index of total factor productivity appropriate to the resolution of the ques-tion of the efficiency of slave labor as that question actually evolved in historical literature?

Much of chapters 5 and 6 of *Time on the Cross* was devoted to reviewing both the pre- and post-Civil War debates on the in-efficiency of slave agriculture and slave labor. From this review it is clear that what the critics of slavery meant was: other in-puts held constant, but substituting slave for free labor and slave managers for free managers, the output of slave farms would be much less than the output of free farms. About this "fact" Clay, Helper, Olmsted, and most other antislavery critics had no doubt. This confidence stemmed from the conviction that slavery "degraded" labor, that slavery turned plantation owners into "idlers," that "comparing man with man," slave laborers were less than half as produc-tive as whites, that Africans were "far less adapted for steady, uninterrupted labor than we are," and that "white laborers of equal intelligence and under equal stimulus will cut twice as much wood, split twice as many rails, and hoe a third more corn a day than Negroes." (See Clay, p. 204 and Olm-sted, pp. 91, 467-68.)

Nor does the geometric index of total fac-tor productivity do violence to the issue of efficiency as it was perceived and discussed by the principal scholars who preceded us. Certainly neither Ulrich Bonnell Phillips, nor Gray, nor Ralph Betts Flanders, nor Robert R. Russel, nor Kenneth M. Stampp were talking about a morally weighted mea-sure of productivity, but of the comparative efficiency of slave and free labor, in just the manner that the issue was raised by the antebellum critics of slavery.

To the extent that the argument for moral weighting is really an objection to using observed prices for aggregating southern output, one needs to assess the effect of slavery on the observed prices. On all agri-cultural commodities except slave-produced staples, the South was a price taker. Since

the South contributed about three-quarters of the world's supply of cotton, its behavior did affect the world price of cotton. In the absence of slavery, however, the supply of cotton would presumably have shifted to the left. This implies that the observed price of cotton was lower, not higher, than it would have been if slaves had been free to make their own choices about the provision of their labor. Consequently, the use of observed prices to aggregate output yields a lower value of G_s/G_n than would be obtained by aggregation based on the counterfactual price of cotton.

Of course, the fact that blacks who toiled on large plantations were more efficient than free workers does not imply that blacks were inherently superior to whites as workers. It was the system that forced men to work at the pace of an assembly line (called the gang) that made slave laborers more efficient than free laborers. Moreover, the gang system, as already noted, appears to have raised productivity only on farms that specialized in certain crops.[31] It should, of course, be emphasized that greater efficiency does not mean greater good. As we attempted to demonstrate in *Time on the Cross*, freedom has value and the loss of freedom by slaves was greater than the gain in measured output to free persons.

[31] It is important to distinguish between the technological characteristics of the production process and the manner in which the labor employed in that process was obtained. While it was force, not volunteerism, that ultimately permitted gang labor to exist, it does not follow that force alone would have led to high levels of productivity, if that potential was not inherent in the production process. If force alone created high levels of productivity, small plantations with 1-15 slaves should have been more efficient than free farms, even though they did not utilize the gang system. Similarly, the argument that some mix of coercion and volunteerism may have been necessary in the initial creation of a factory labor force does not rule out economies of scale or other technological efficiencies as features of the factory system.

REFERENCES

R. V. Anderson, "Labor Utilization and Productivity, Diversification and Self Sufficiency, Southern Plantations, 1800-1840," unpublished doctoral dissertation, Univ. No. Carolina 1974.

Thomas S. Berry, *Western Prices Before 1861* Cambridge, Mass. 1943.

Percy W. Bidwell and John I. Falconer, *History of Agriculture in the Northern United States, 1620-1860*, Washington 1925.

John E. Cairnes, *The Slave Power: Its Character, Career, and Probable Designs: Being an Attempt to Explain the Real Issues Involved in the American Contest*, introduction by Harold D. Woodman, New York 1969.

Cassius M. Clay, *The Writings of Cassius Marcellus Clay: Including Speeches and Addresses*, New York 1848.

Arthur H. Cole, *Wholesale Commodity Prices in the United States, Statistical Supplement*, Cambridge, Mass. 1938.

J. R. Covert, *Seedtime and Harvest*, U.S.D.A. Bur. Statist., Bull. 85, Washington 1912.

Clarence H. Danhof, *Change in Agriculture: The Northern United States, 1820-1870*, Cambridge, Mass. 1969.

Edwin A. Davis, *Plantation Life in the Florida Parishes of Louisiana 1836-1844, as Reflected in the Diary of Bennet H. Barrow*, New York 1943.

L. E. Davis, "One Potato, Two Potato, Sweet Potato Pie: Clio Looks at Slavery and the South," paper presented to the MSSB-Univ. Rochester conference on *Time on the Cross*, 1974.

P. A. David and P. Temin, "Slavery: The Progressive Institution?" *J. Econ. Hist.*, Sept. 1974, *34*, 739-83.

E. D. Domar, "On the Measurement of Technological Change," *Econ. J.*, Dec. 1961, *71*, 709-29.

Ralph B. Flanders, *Plantation Slavery in Georgia*, Chapel Hill 1933.

Robert W. Fogel, *Railroads and American Economic Growth: Essays in Econometric History*, Baltimore 1964.

———, "American Interregional Trade in the Nineteenth Century," in Ralph Andreano, ed., *New Views on American Economic Development: A Selective Anthology of Recent Work*, Cambridge, Mass. 1965.

——— **and S. L. Engerman,** "The Relative Efficiency of Slavery: A Comparison of Northern and Southern Agriculture in

1860," *Explor. Econ. Hist.*, Spring 1971, *8*, 353-67.

——— and ———, (1974a) *Time on the Cross*, Vols. I, II, Boston 1974.

——— and ———, (1974b) "Further Evidence on the Nutritional Adequacy of the Slave Diet," Univ. Rochester 1974.

——— and ———, "The Relative Efficiency of Slave and Free Agriculture in 1860 and 1850," Harvard Univ. 1975.

——— and ———, *Further Evidence on the Economics of American Negro Slavery*, forthcoming.

J. D. Foust, "The Yeoman Farmer and Westward Expansion of U.S. Cotton Production," unpublished doctoral dissertation, Univ. No. Carolina 1967.

R. E. Gallman, "The Agricultural Sector and the Pace of Economic Growth: U.S. Experience in the Nineteenth Century," in David C. Klingaman and Richard K. Vedder, eds., *Essays in Nineteenth Century Economic History: The Old Northwest*, Athens, Ohio 1975.

Paul W. Gates, *The Farmer's Age: Agriculture 1815-1860*, New York 1960.

Eugene D. Genovese, *The Political Economy of Slavery: Studies in the Economy and Society of the Slave South*, New York 1965.

Claudia D. Goldin, *Urban Slavery in the American South, 1820-1860: A Quantitative History*, Chicago 1976.

Lewis C. Gray, *History of Agriculture in the Southern United States to 1860*, 2 Vols., Washington 1933.

T. H. Haskell, "Were Slaves More Efficient: Some Doubts About *Time on the Cross*," *New York Rev. of Books*, Sept. 19, 1974, 38-42.

———, "The True and Tragical History of *Time on the Cross*," *New York Rev. of Books*, Oct. 2, 1975, 33-39.

Hinton R. Helper, *The Impending Crisis of the South: How to Meet It*, Cambridge, Mass. 1968.

J. A. Hopkins, *Changing Technology and Employment in Agriculture*, U.S. Bur. Agr. Econ., Washington 1941.

S. Lebergott, "Labor Force and Employment, 1800-1960," in Dorothy S. Brady, ed., *Output, Employment, and Productivity in the United States After 1800*, Nat. Bur. Econ. Res., *Stud. in Income and Wealth*, Vol. 30, New York 1966.

J. Metzer, "Rational Management, Modern Business Practices, and Economies of Scale in the Ante-Bellum Southern Plantations," *Explor. Econ. Hist.*, Apr. 1975, *12*, 123-50.

Frederick L. Olmsted, *The Cotton Kingdom*, New York 1953.

J. R. Olson, "Clock-Time vs. Real-Time: A Comparison of the Lengths of the Northern and Southern Agricultural Work-Years," mimeo., Univ. Conn. 1976.

Ulrich B. Phillips, *American Negro Slavery: A Survey of the Supply, Employment and Control of Negro Labor as Determined by the Plantation Regime*, New York 1918.

M. L. Primack, "Farm Formed Capital in American Agriculture: 1850 to 1910," unpublished doctoral dissertation, Univ. No. Carolina 1962.

R. R. Russel, "The General Effects of Slavery upon Southern Economic Progress," *J. Southern Hist.*, Feb. 1938, *4*, 34-54.

Robert Russell, *North America: Its Agriculture and Climate*, Edinburgh 1857.

Kenneth M. Stampp, *The Peculiar Institution: Slavery in the Ante-Bellum South*, New York 1956.

M. W. Towne and W. D. Rasmussen, "Farm Gross Product and Gross Investment in the Nineteenth Century," in *Trends in the American Economy in the Nineteenth Century*, Nat. Bur. Econ. Res., *Stud. in Income and Wealth*, Vol. 24, Princeton 1960.

J. L. Watkins, *Production and Price of Cotton for One Hundred Years*, USDA Misc. Series, Bull. No. 9, Washington 1895.

H. Woodman, "The Old South and the New History," paper presented to the MSSB-Univ. Rochester conference on *Time on the Cross*, 1974.

G. Wright, "The Economics of Cotton in the Antebellum South," unpublished doctoral dissertation, Yale Univ. 1969.

———, "An Econometric Study of Cotton Production and Trade, 1830-1860," *Rev. Econ. Statist.*, May 1971, *53*, 111-20.

———, "The Economic Analysis of *Time on the Cross*," paper presented to the

MSSB-Univ. Rochester conference on *Time on the Cross*, 1974.

———, "Slavery and the Cotton Boom," *Explor. Econ. Hist.*, Oct. 1975, *12*, 439–51.

U.S. Bureau of the Census, *Preliminary Report of the Eighth Census, 1860*, Washington 1862.

———, *Eleventh Census of the United States: 1890; Report on the Statistics of Agriculture in the United States*, Washington 1895.

———, *Historical Statistics of the United States, Colonial Times to 1957*, Washington 1960.

U.S. Department of Agriculture, *Report of the Commissioner of Agriculture, 1867*, Washington 1868.

———, Agr. Marketing Service, *Corn Acreage, Yield, and Production, 1866–1943*, Washington 1954.

———, Agr. Marketing Service, *Cotton and Cottonseed*, Statist. Bull. 164, Washington 1955.

Monopolistic Competition and Optimum Product Diversity

By AVINASH K. DIXIT AND JOSEPH E. STIGLITZ*

The basic issue concerning production in welfare economics is whether a market solution will yield the socially optimum kinds and quantities of commodities. It is well known that problems can arise for three broad reasons: distributive justice; external effects; and scale economies. This paper is concerned with the last of these.

The basic principle is easily stated.[1] A commodity should be produced if the costs can be covered by the sum of revenues and a properly defined measure of consumer's surplus. The optimum amount is then found by equating the demand price and the marginal cost. Such an optimum can be realized in a market if perfectly discriminatory pricing is possible. Otherwise we face conflicting problems. A competitive market fulfilling the marginal condition would be unsustainable because total profits would be negative. An element of monopoly would allow positive profits, but would violate the marginal condition.[2] Thus we expect a market solution to be suboptimal. However, a much more precise structure must be put on the problem if we are to understand the nature of the bias involved.

It is useful to think of the question as one of quantity versus diversity. With scale economies, resources can be saved by producing fewer goods and larger quantities of each. However, this leaves less variety, which entails some welfare loss. It is easy and probably not too unrealistic to model scale economies by supposing that each potential commodity involves some fixed set-up cost and has a constant marginal cost. Modeling the desirability of variety has been thought to be difficult, and several indirect approaches have been adopted. The Hotelling spatial model, Lancaster's product characteristics approach, and the mean-variance portfolio selection model have all been put to use.[3] These lead to results involving transport costs or correlations among commodities or securities, and are hard to interpret in general terms. We therefore take a direct route, noting that the convexity of indifference surfaces of a conventional utility function defined over the quantities of all potential commodities already embodies the desirability of variety. Thus, a consumer who is indifferent between the quantities (1,0) and (0,1) of two commodities prefers the mix (1/2,1/2) to either extreme. The advantage of this view is that the results involve the familiar own- and cross-elasticities of demand functions, and are therefore easier to comprehend.

There is one case of particular interest on which we concentrate. This is where potential commodities in a group or sector or industry are good substitutes among themselves, but poor substitutes for the other commodities in the economy. Then we are led to examining the market solution in relation to an optimum, both as regards biases within the group, and between the group and the rest of the economy. We expect the answer to depend on the intra- and intersector elasticities of substitution. To demonstrate the point as simply as possible, we shall aggregate the rest of the economy into one good labeled 0, chosen as the numeraire. The economy's endowment of it is normalized at unity; it can be thought of as the time at the disposal of the consumers.

*Professors of economics, University of Warwick and Stanford University, respectively. Stiglitz's research was supported in part by NSF Grant SOC74-22182 at the Institute for Mathematical Studies in the Social Sciences, Stanford. We are indebted to Michael Spence, to a referee, and the managing editor for comments and suggestions on earlier drafts.

[1] See also the exposition by Michael Spence.

[2] A simple exposition is given by Peter Diamond and Daniel McFadden.

[3] See the articles by Harold Hotelling, Nicholas Stern, Kelvin Lancaster, and Stiglitz.

The potential range of related products is labeled 1,2,3,.... Writing the amounts of the various commodities as x_0 and $x = (x_1, x_2, x_3 \ldots$, we assume a separable utility function with convex indifference surfaces:

$$u = U(x_0, V(x_1, x_2, x_3 \ldots)) \tag{1}$$

In Sections I and II we simplify further by assuming that V is a symmetric function, and that all commodities in the group have equal fixed and marginal costs. Then the actual labels given to commodities are immaterial, even though the total number n being produced is relevant. We can thus label these commodities $1, 2, \ldots, n$, where the potential products $(n+1), (n+2), \ldots$ are not being produced. This is a restrictive assumption, for in such problems we often have a natural asymmetry owing to graduated physical differences in commodities, with a pair close together being better mutual substitutes than a pair farther apart. However, even the symmetric case yields some interesting results. In Section III, we consider some aspects of asymmetry.

We also assume that all commodities have unit income elasticities. This differs from a similar recent formulation by Michael Spence, who assumes U linear in x_0, so that the industry is amenable to partial equilibrium analysis. Our approach allows a better treatment of the intersectoral substitution, but the other results are very similar to those of Spence.

We consider two special cases of (1). In Section I, V is given a *CES* form, but U is allowed to be arbitrary. In Section II, U is taken to be Cobb-Douglas, but V has a more general additive form. Thus the former allows more general intersector relations, and the latter more general intrasector substitution, highlighting different results.

Income distribution problems are neglected. Thus U can be regarded as representing Samuelsonian social indifference curves, or (assuming the appropriate aggregation conditions to be fulfilled) as a multiple of a representative consumer's utility. Product diversity can then be interpreted either as different consumers using different varieties, or as diversification on the part of each consumer.

I. Constant-Elasticity Case

A. *Demand Functions*

The utility function in this section is

$$u = U\left(x_0, \left\{\sum_i x_i^\rho\right\}^{1/\rho}\right) \tag{2}$$

For concavity, we need $\rho < 1$. Further, since we want to allow a situation where several of the x_i are zero, we need $\rho > 0$. We also assume U homothetic in its arguments.

The budget constraint is

$$x_0 + \sum_{i=1}^{n} p_i x_i = I \tag{3}$$

where p_i are prices of the goods being produced, and I is income in terms of the numeraire, i.e., the endowment which has been set at 1 plus the profits of the firms distributed to the consumers, or minus the lump sum deductions to cover the losses, as the case may be.

In this case, a two-stage budgeting procedure is valid.[4] Thus we define dual quantity and price indices

$$y = \left\{\sum_{i=1}^{n} x_i^\rho\right\}^{1/\rho} \qquad q = \left\{\sum_{i=1}^{n} p_i^{-1/\beta}\right\}^{-\beta} \tag{4}$$

where $\beta = (1-\rho)/\rho$, which is positive since $0 < \rho < 1$. Then it can be shown[5] that in the first stage,

$$y = I\frac{s(q)}{q} \qquad x_0 = I(1 - s(q)) \tag{5}$$

for a function s which depends on the form of U. Writing $\sigma(q)$ for the elasticity of substitution between x_0 and y, we define $\theta(q)$ as the elasticity of the function s, i.e., $qs'(q)/s(q)$. Then we find

$$\theta(q) = \{1 - \sigma(q)\}\{1 - s(q)\} < 1 \tag{6}$$

but $\theta(q)$ can be negative as $\sigma(q)$ can exceed 1.

[4]See p. 21 of John Green.

[5]These details and several others are omitted to save space, but can be found in the working paper by the authors, cited in the references.

Turning to the second stage of the problem, it is easy to show that for each i,

$$x_i = y\left[\frac{q}{p_i}\right]^{1/(1-\rho)} \tag{7}$$

where y is defined by (4). Consider the effect of a change in p_i alone. This affects x_i directly, and also through q; thence through y as well. Now from (4) we have the elasticity

$$\frac{\partial \log q}{\partial \log p_i} = \left(\frac{q}{p_i}\right)^{1/\beta} \tag{8}$$

So long as the prices of the products in the group are not of different orders of magnitude, this is of the order $(1/n)$. We shall assume that n is reasonably large, and accordingly neglect the effect of each p_i on q; thus the indirect effects on x_i. This leaves us with the elasticity

$$\frac{\partial \log x_i}{\partial \log p_i} = \frac{-1}{(1-\rho)} = \frac{-(1+\beta)}{\beta} \tag{9}$$

In the Chamberlinian terminology, this is the elasticity of the dd curve, i.e., the curve relating the demand for each product type to its own price with all other prices held constant.

In our large group case, we also see that for $i \neq j$, the cross elasticity $\partial \log x_i/\partial \log p_j$ is negligible. However, if all prices in the group move together, the individually small effects add to a significant amount. This corresponds to the Chamberlinian DD curve. Consider a symmetric situation where $x_i = x$ and $p_i = p$ for all i from 1 to n. We have

$$\begin{aligned} y &= xn^{1/\rho} = xn^{1+\beta} \\ q &= pn^{-\beta} = pn^{-(1-\rho)/\rho} \end{aligned} \tag{10}$$

and then from (5) and (7),

$$x = \frac{Is(q)}{pn} \tag{11}$$

The elasticity of this is easy to calculate; we find

$$\frac{\partial \log x}{\partial \log p} = -[1 - \theta(q)] \tag{12}$$

Then (6) shows that the DD curve slopes downward. The conventional condition that the dd curve be more elastic is seen from (9) and (12) to be

$$\frac{1}{\beta} + \theta(q) > 0 \tag{13}$$

Finally, we observe that for $i \neq j$,

$$\frac{x_i}{x_j} = \left[\frac{p_j}{p_i}\right]^{1/(1-\rho)} \tag{14}$$

Thus $1/(1-\rho)$ is the elasticity of substitution between any two products within the group.

B. *Market Equilibrium*

It can be shown that each commodity is produced by one firm. Each firm attempts to maximize its profit, and entry occurs until the marginal firm can only just break even. Thus our market equilibrium is the familiar case of Chamberlinian monopolistic competition, where the question of quantity versus diversity has often been raised.[6] Previous analyses have failed to consider the desirability of variety in an explicit form, and have neglected various intra- and intersector interactions in demand. As a result, much vague presumption that such an equilibrium involves excessive diversity has built up at the back of the minds of many economists. Our analysis will challenge several of these ideas.

The profit-maximization condition for each firm acting on its own is the familiar equality of marginal revenue and marginal cost. Writing c for the common marginal cost, and noting that the elasticity of demand for each firm is $(1+\beta)/\beta$, we have for each active firm:

$$p_i\left(1 - \frac{\beta}{1+\beta}\right) = c$$

Writing p_e for the common equilibrium price for each variety being produced, we have

$$p_e = c(1+\beta) = \frac{c}{\rho} \tag{15}$$

[6]See Edwin Chamberlin, Nicholas Kaldor, and Robert Bishop.

The second condition for equilibrium is that firms enter until the next potential entrant would make a loss. If n is large enough so that 1 is a small increment, we can assume that the marginal firm is exactly breaking even, i.e., $(p_n - c)x_n = a$, where x_n is obtained from the demand function and a is the fixed cost. With symmetry, this implies zero profit for all intramarginal firms as well. Then $I = 1$, and using (11) and (15) we can write the condition so as to yield the number n_e of active firms:

$$\frac{s(p_e n_e^{-\beta})}{p_e n_e} = \frac{a}{\beta c} \tag{16}$$

Equilibrium is unique provided $s(p_e n^{-\beta})/p_e n$ is a monotonic function of n. This relates to our earlier discussion about the two demand curves. From (11) we see that the behavior of $s(pn^{-\beta})/pn$ as n increases tells us how the demand curve DD for each firm shifts as the number of firms increases. It is natural to assume that it shifts to the left, i.e., the function above decreases as n increases for each fixed p. The condition for this in elasticity form is easily seen to be

$$1 + \beta\theta(q) > 0 \tag{17}$$

This is exactly the same as (13), the condition for the dd curve to be more elastic than the DD curve, and we shall assume that it holds.

The condition can be violated if $\sigma(q)$ is sufficiently higher than one. In this case, an increase in n lowers q, and shifts demand towards the monopolistic sector to such an extent that the demand curve for each firm shifts to the right. However, this is rather implausible.

Conventional Chamberlinian analysis assumes a fixed demand curve for the group as a whole. This amounts to assuming that $n \cdot x$ is independent of n, i.e., that $s(pn^{-\beta})$ is independent of n. This will be so if $\beta = 0$, or if $\sigma(q) = 1$ for all q. The former is equivalent to assuming that $\rho = 1$, when all products in the group are perfect substitutes, i.e., diversity is not valued at all. That would be contrary to the intent of the whole analysis. Thus, implicitly, conventional analysis assumes $\sigma(q) = 1$. This gives a constant budget share for the monopolistically competitive sector. Note that in our parametric formulation, this implies a unit-elastic DD curve, (17) holds, and so equilibrium is unique.

Finally, using (7), (11), and (16), we can calculate the equilibrium output for each active firm:

$$x_e = \frac{a}{\beta c} \tag{18}$$

We can also write an expression for the budget share of the group as a whole:

$$s_e = s(q_e) \tag{19}$$

where $$q_e = p_e n_e^{-\beta}$$

These will be useful for subsequent comparisons.

C. *Constrained Optimum*

The next task is to compare the equilibrium with a social optimum. With economies of scale, the first best or unconstrained (really constrained only by technology and resource availability) optimum requires pricing below average cost, and therefore lump sum transfers to firms to cover losses. The conceptual and practical difficulties of doing so are clearly formidable. It would therefore appear that a more appropriate notion of optimality is a constrained one, where each firm must have nonnegative profits. This may be achieved by regulation, or by excise or franchise taxes or subsidies. The important restriction is that lump sum subsidies are not available.

We begin with such a constrained optimum. The aim is to choose n, p_i, and x_i so as to maximize utility, satisfying the demand functions and keeping the profit for each firm nonnegative. The problem is somewhat simplified by the result that all active firms should have the same output levels and prices, and should make exactly zero profit. We omit the proof. Then we can set $I = 1$, and use (5) to express utility as a function of q alone. This is of course a decreasing function. Thus the problem of maximizing u becomes that of minimizing q, i.e.,

$$\min_{n,p} pn^{-\beta}$$

subject to

$$(20)\qquad (p - c)\frac{s(pn^{-\beta})}{pn} = a$$

To solve this, we calculate the logarithmic marginal rate of substitution along a level curve of the objective, the similar rate of transformation along the constraint, and equate the two. This yields the condition

$$(21)\qquad \frac{\frac{c}{p-c} + \theta(q)}{1 + \beta\theta(q)} = \frac{1}{\beta}$$

The second-order condition can be shown to hold, and (21) simplifies to yield the price for each commodity produced in the constrained optimum, p_c, as

$$(22)\qquad p_c = c(1 + \beta)$$

Comparing (15) and (22), we see that the two solutions have the same price. Since they face the same break-even constraint, they have the same number of firms as well, and the values for all other variables can be calculated from these two. Thus we have a rather surprising case where the monopolistic competition equilibrium is identical with the optimum constrained by the lack of lump sum subsidies. Chamberlin once suggested that such an equilibrium was "a sort of ideal"; our analysis shows when and in what sense this can be true.

D. *Unconstrained Optimum*

These solutions can in turn be compared to the unconstrained or first best optimum. Considerations of convexity again establish that all active firms should produce the same output. Thus we are to choose n firms each producing output x in order to maximize

$$(23)\qquad u = U(1 - n(a + cx), xn^{1+\beta})$$

where we have used the economy's resource balance condition and (10). The first-order conditions are

$$(24)\qquad -ncU_0 + n^{1+\beta}U_y = 0$$

$$(25)\qquad -(a + cx)U_0 + (1 + \beta)xn^{\beta}U_y = 0$$

From the first stage of the budgeting problem, we know that $q = U_y/U_0$. Using (24) and (10), we find the price charged by each active firm in the unconstrained optimum, p_u, equal to marginal cost

$$(26)\qquad p_u = c$$

This, of course, is no surprise. Also from the first-order conditions, we have

$$(27)\qquad x_u = \frac{a}{c\beta}$$

Finally, with (26), each active firm covers its variable cost exactly. The lump sum transfers to firms then equal an, and therefore $I = 1 - an$, and

$$x = (1 - an)\frac{s(pn^{-\beta})}{pn}$$

The number of firms n_u is then defined by

$$(28)\qquad \frac{s(cn_u^{-\beta})}{n_u} = \frac{a/\beta}{1 - an_u}$$

We can now compare these magnitudes with the corresponding ones in the equilibrium or the constrained optimum. The most remarkable result is that the output of each active firm is the same in the two situations. The fact that in a Chamberlinian equilibrium each firm operates to the left of the point of minimum average cost has been conventionally described by saying that there is excess capacity. However, when variety is desirable, i.e., when the different products are not perfect substitutes, it is not in general optimum to push the output of each firm to the point where all economies of scale are exhausted.[7] We have shown in one case that is not an extreme one, that the first best optimum does not exploit economies of scale beyond the extent achieved in the equilibrium. We can then easily conceive of cases where the equilibrium exploits economies of scale too far from the point of view of social optimality. Thus our results undermine the validity of the folklore of excess capacity, from the point of view of the

[7]See David Starrett.

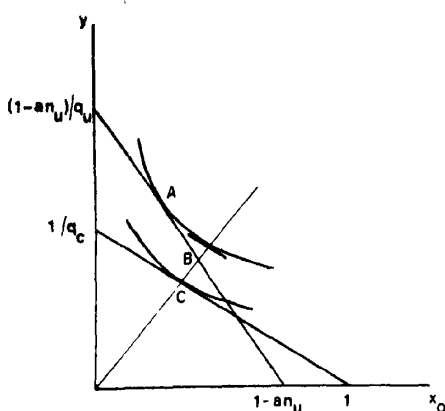



FIGURE 1

unconstrained optimum as well as the constrained one.

A direct comparison of the numbers of firms from (16) and (28) would be difficult, but an indirect argument turns out to be simple. It is clear that the unconstrained optimum has higher utility than the constrained optimum. Also, the level of lump sum income in it is less than that in the latter. It must therefore be the case that

$$q_u < q_c = q_e \tag{29}$$

Further, the difference must be large enough that the budget constraint for x_0 and the quantity index y in the unconstrained case must lie outside that in the constrained case in the relevant region, as shown in Figure 1. Let C be the constrained optimum, A the unconstrained optimum, and let B be the point where the line joining the origin to C meets the indifference curve in the unconstrained case. By homotheticity the indifference curve at B is parallel to that at C, so each of the moves from C to B and from B to A increases the value of y. Since the value of x is the same in the two optima, we must have

$$n_u > n_c = n_e \tag{30}$$

Thus the unconstrained optimum actually allows more variety than the constrained optimum and the equilibrium; this is another point contradicting the folklore on excessive diversity.

Using (29) we can easily compare the budget shares. In the notation we have been using, we find $s_u \gtrless s_c$ as $\theta(q) \gtrless 0$, i.e., as $\sigma(q) \lessgtr 1$ providing these hold over the entire relevant range of q.

It is not possible to have a general result concerning the relative magnitudes of x_0 in the two situations; an inspection of Figure 1 shows this. However, we have a sufficient condition:

$$x_{0u} = (1 - an_u)(1 - s_u) < 1 - s_u \leq 1 - s_c$$
$$= x_{0c} \text{ if } \sigma(q) \geq 1$$

In this case the equilibrium or the constrained optimum use more of the numeraire resource than the unconstrained optimum. On the other hand, if $\sigma(q) = 0$ we have L-shaped isoquants, and in Figure 1, points A and B coincide giving the opposite conclusion.

In this section we have seen that with a constant intrasector elasticity of substitution, the market equilibrium coincides with the constrained optimum. We have also shown that the unconstrained optimum has a greater number of firms, each of the same size. Finally, the resource allocation between the sectors is shown to depend on the intersector elasticity of substitution. This elasticity also governs conditions for uniqueness of equilibrium and the second-order conditions for an optimum.

Henceforth we will achieve some analytic simplicity by making a particular assumption about intersector substitution. In return, we will allow a more general form of intrasector substitution.

II. Variable Elasticity Case

The utility function is now

$$u = x_0^{1-\gamma}\left\{\sum_i v(x_i)\right\}^\gamma \tag{31}$$

with v increasing and concave, $0 < \gamma < 1$. This is somewhat like assuming a unit intersector elasticity of substitution. However, this is not rigorous since the group utility $V(\underline{x}) = \sum_i v(x_i)$ is not homothetic and therefore two-stage budgeting is not applicable.

It can be shown that the elasticity of the dd curve in the large group case is

$$(32)\quad -\frac{\partial \log x_i}{\partial \log p_i} = -\frac{v'(x_i)}{x_i v''(x_i)} \quad \text{for any } i$$

This differs from the case of Section I in being a function of x_i. To highlight the similarities and the differences, we define $\beta(x)$ by

$$(33)\quad \frac{1+\beta(x)}{\beta(x)} = -\frac{v'(x)}{xv''(x)}$$

Next, setting $x_i = x$ and $p_i = p$ for $i = 1, 2, \ldots, n$, we can write the DD curve and the demand for the numeraire as

$$(34)\quad x = \frac{I}{np}\omega(x), \qquad x_0 = I[1-\omega(x)]$$

where

$$(35)\quad \omega(x) = \frac{\gamma\rho(x)}{[\gamma\rho(x) + (1-\gamma)]}$$

$$\rho(x) = \frac{xv'(x)}{v(x)}$$

We assume that $0 < \rho(x) < 1$, and therefore have $0 < \omega(x) < 1$.

Now consider the Chamberlinian equilibrium. The profit-maximization condition for each active firm yields the common equilibrium price p_e in terms of the common equilibrium output x_e as

$$(36)\quad p_e = c[1+\beta(x_e)]$$

Note the analogy with (15). Substituting (36) in the zero pure profit condition, we have x_e defined by

$$(37)\quad \frac{cx_e}{a+cx_e} = \frac{1}{1+\beta(x_e)}$$

Finally, the number of firms can be calculated using the DD curve and the break-even condition, as

$$(38)\quad n_e = \frac{\omega(x_e)}{a+cx_e}$$

For uniqueness of equilibrium we once again use the conditions that the dd curve is more elastic than the DD curve, and that entry shifts the DD curve to the left. However, these conditions are rather involved and opaque, so we omit them.

Let us turn to the constrained optimum. We wish to choose n and x to maximize u, subject to (34) and the break-even condition $px = a + cx$. Substituting, we can express u as a function of x alone:

$$(39)\quad u = \gamma^{\gamma}(1-\gamma)^{(1-\gamma)} \frac{\left[\frac{\rho(x)v(x)}{a+cx}\right]^{\gamma}}{\gamma\rho(x) + (1-\gamma)}$$

The first-order condition defines x_c:

$$(40)\quad \frac{cx_c}{a+cx_c} = \frac{1}{1+\beta(x_c)} - \frac{\omega(x_c)x_c\rho'(x_c)}{\gamma\rho(x_c)}$$

Comparing this with (37) and using the second-order condition, it can be shown that provided $\rho'(x)$ is one-signed for all x,

$$(41)\quad x_c \gtrless x_e \text{ according as } \rho'(x) \lesseqgtr 0$$

With zero pure profit in each case, the points (x_e, p_e) and (x_c, p_c) lie on the same declining average cost curve, and therefore

$$(42)\quad p_c \lesseqgtr p_e \text{ according as } x_c \gtrless x_e$$

Next we note that the dd curve is tangent to the average cost curve at (x_e, p_e) and the DD curve is steeper. Consider the case $x_c > x_e$. Now the point (x_c, p_c) must lie on a DD curve further to the right than (x_e, p_e), and therefore must correspond to a smaller number of firms. The opposite happens if $x_c < x_e$. Thus,

$$(43)\quad n_c \lesseqgtr n_e \text{ according as } x_c \gtrless x_e$$

Finally, (41) shows that in both cases that arise there, $\rho(x_c) < \rho(x_e)$. Then $\omega(x_c) < \omega(x_e)$, and from (34),

$$(44)\quad x_{0c} > x_{0e}$$

A smaller degree of intersectoral substitution could have reversed the result, as in Section I.

An intuitive reason for these results can be given as follows. With our large group assumptions, the revenue of each firm is proportional to $xv'(x)$. However, the contribution of its output to group utility is $v(x)$. The ratio of the two is $\rho(x)$. Therefore, if $\rho'(x) > 0$, then at the margin each firm finds it more profitable to expand than what would be socially desirable, so $x_e > x_c$.

Given the break-even constraint, this leads to there being fewer firms.

Note that the relevant magnitude is the elasticity of utility, and not the elasticity of demand. The two are related, since

$$(45) \quad x\frac{\rho'(x)}{\rho(x)} = \frac{1}{1+\beta(x)} - \rho(x)$$

Thus, if $\rho(x)$ is constant over an interval, so is $\beta(x)$ and we have $1/(1+\beta) = \rho$, which is the case of Section I. However, if $\rho(x)$ varies, we cannot infer a relation between the signs of $\rho'(x)$ and $\beta'(x)$. Thus the variation in the elasticity of demand is not in general the relevant consideration. However, for important families of utility functions there is a relationship. For example, for $v(x) = (k + mx)^j$, with $m > 0$ and $0 < j < 1$, we find that $-xv''/v'$ and xv'/v are positively related. Now we would normally expect that as the number of commodities produced increases, the elasticity of substitution between any pair of them should increase. In the symmetric equilibrium, this is just the inverse of the elasticity of marginal utility. Then a higher x would correspond to a lower n, and therefore a lower elasticity of substitution, higher $-xv''/v'$ and higher xv'/v. Thus we are led to expect that $\rho'(x) > 0$, i.e., that the equilibrium involves fewer and bigger firms than the constrained optimum. Once again the common view concerning excess capacity and excessive diversity in monopolistic competition is called into question.

The unconstrained optimum problem is to choose n and x to maximize

$$(46) \quad u = [nv(x)]^{\gamma}[1 - n(a + cx)]^{1-\gamma}$$

It is easy to show that the solution has

$$(47) \quad p_u = c$$

$$(48) \quad \frac{cx_u}{a + cx_u} = \rho(x_u)$$

$$(49) \quad n_u = \frac{\gamma}{a + cx_u}$$

Then we can use the second-order condition to show that

$$(50) \quad x_u \lesseqgtr x_c \text{ according as } \rho'(x) \gtreqless 0$$

This is in each case transitive with (41), and therefore yields similar output comparisons between the equilibrium and the unconstrained optimum.

The price in the unconstrained optimum is of course the lowest of the three. As to the number of firms, we note

$$n_c = \frac{\omega(x_c)}{a + cx_c} < \frac{\gamma}{a + cx_c}$$

and therefore we have a one-way comparison:

$$(51) \quad \text{If } x_u < x_c, \text{ then } n_u > n_c$$

Similarly for the equilibrium. These leave open the possibility that the unconstrained optimum has both bigger and more firms. That is not unreasonable; after all the unconstrained optimum uses resources more efficiently.

III. Asymmetric Cases

The discussion so far imposed symmetry within the group. Thus the number of varieties being produced was relevant, but any group of n was just as good as any other group of n. The next important modification is to remove this restriction. It is easy to see how interrelations within the group of commodities can lead to biases. Thus, if no sugar is being produced, the demand for coffee may be so low as to make its production unprofitable when there are set-up costs. However, this is open to the objection that with complementary commodities, there is an incentive for one entrant to produce both. However, problems exist even when all the commodities are substitutes. We illustrate this by considering an industry which will produce commodities from one of two groups, and examine whether the choice of the wrong group is possible.[8]

Suppose there are two sets of commodities beside the numeraire, the two being perfect substitutes for each other and each having a constant elasticity subutility function. Further, we assume a constant budget share

[8]For an alternative approach using partial equilibrium methods, see Spence.

r the numeraire. Thus the utility function

2)

$$\iota = x_0^{1-s}\left\{\left[\sum_{i_1=1}^{n} x_{i_1}^{\rho_1}\right]^{1/\rho_1} + \left[\sum_{i_2=1}^{n_2} x_i^{\rho_2}\right]^{1/\rho_2}\right\}^s$$

'e assume that each firm in group i has a ‹ed cost a_i and a constant marginal cost c_i. Consider two types of equilibria, only ıe commodity group being produced in ıch. These are given by

3a) $$\bar{x}_1 = \frac{a_1}{c_1\beta_1}, \bar{x}_2 = 0$$

$$\bar{p}_1 = c_1(1 + \beta_1)$$

$$\bar{n}_1 = \frac{s\beta_1}{a_1(1 + \beta_1)}$$

$$\bar{q}_1 = \bar{p}_1\bar{n}_1^{-\beta_1} = c_1(1 + \beta_1)^{1+\beta_1}\left(\frac{a_1}{s}\right)^{\beta_1}$$

$$\bar{u}_1 = s^s(1 - s)^{1-s}\bar{q}_1^{-s}$$

3b) $$\bar{x}_2 = \frac{a_2}{c_2\beta_2}, \bar{x}_1 = 0$$

$$\bar{p}_2 = c_2(1 + \beta_2)$$

$$\bar{n}_2 = \frac{s\beta_2}{a_2(1 + \beta_2)}$$

$$\bar{q}_2 = \bar{p}_2\bar{n}_2^{-\beta_2} = c_2(1 + \beta_2)^{1+\beta_2}\left(\frac{a_2}{s}\right)^{\beta_2}$$

$$\bar{u}_2 = s^s(1 - s)^{1-s}\bar{q}_2^{-s}$$

Equation (53a) is a Nash equilibrium if d only if it does not pay a firm to produce commodity of the second group. The de-and for such a commodity is

$$x_2 = \begin{cases} 0 & \text{for } p_2 \geq \bar{q}_1 \\ s/p_2 & \text{for } p_2 < \bar{q}_1 \end{cases}$$

ence we require

$$\max_{p_2} (p_2 - c_2)x_2 = s\left(1 - \frac{c_2}{\bar{q}_1}\right) < a_2$$

4) $$\bar{q}_1 < \frac{sc_2}{s - a_2}$$

milarly, (53b) is a Nash equilibrium if and only if

(55) $$\bar{q}_2 < \frac{sc_1}{s - a_1}$$

Now consider the optimum. Both the objective and the constraint are such as to lead the optimum to the production of commodities from only one group. Thus, suppose n_i commodities from group i are being produced at levels x_i each, and offered at prices p_i. The utility level is given by

(56) $$u = x_0^{1-s}\{x_1 n_1^{1+\beta_1} + x_2 n_2^{1+\beta_2}\}^s$$

and the resource availability constraint is

(57)
$$x_0 + n_1(a_1 + c_1x_1) + n_2(a_2 + c_2x_2) = 1$$

Given the values of the other variables, the level curves of u in (n_1, n_2) space are concave to the origin, while the constraint is linear. We must therefore have a corner optimum. (As for the break-even constraint, unless the two $q_i = p_i n_i^{-\beta_i}$ are equal, the demand for commodities in one group is zero, and there is no possibility of avoiding a loss there.)

Note that we have structured our example so that if the correct group is chosen, the equilibrium will not introduce any further biases in relation to the constrained optimum. Therefore, to find the constrained optimum, we only have to look at the values of $\bar{u}_i$ in (53a) and (53b) and see which is the greater. In other words, we have to see which $\bar{q}_i$ is the smaller, and choose the situation (which may or may not be a Nash equilibrium) defined in (53a) and (53b) corresponding to it.

Figure 2 is drawn to depict the possible equilibria and optima. Given all the relevant parameters, we calculate $(\bar{q}_1, \bar{q}_2)$ from (53a) and (53b). Then (54) and (55) tell us whether either or both of the situations are possible equilibria, while a simple comparison of the magnitudes of $\bar{q}_1$ and $\bar{q}_2$ tells us which is the constrained optimum. In the figure, the nonnegative quadrant is split into regions in each of which we have one combination of equilibria and optima. We only have to locate the point $(\bar{q}_1, \bar{q}_2)$ in this space to know the result for the given

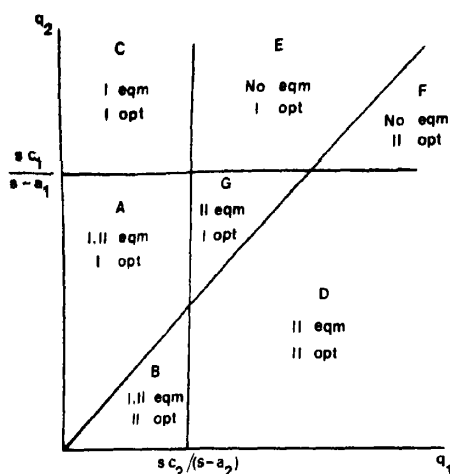



FIGURE 2 SOLUTIONS LABELED I REFER TO EQUATION (53a), SOLUTIONS LABELED II REFER TO EQUATION (53b)

parameter values. Moreover, we can compare the location of the points corresponding to different parameter values and thus do some comparative statics.

To understand the results, we must examine how $\bar{q}_i$ depends on the relevant parameters. It is easy to see that each is an increasing function of a_i and c_i. We also find

$$\frac{\partial \log \bar{q}_i}{\partial \beta_i} = -\log \bar{n}_i \tag{58}$$

and we expect this to be large and negative. Further, we see from (9) that a higher β_i corresponds to a lower own-price elasticity of demand for each commodity in that group. Thus $\bar{q}_i$ is an increasing function of this elasticity.

Consider initially a symmetric situation, with $sc_1/(s - a_1) = sc_2/(s - a_2)$, $\beta_1 = \beta_2$ (the region G vanishes then), and suppose the point $(\bar{q}_1, \bar{q}_2)$ is on the boundary between regions A and B. Now consider a change in one parameter, say, a higher own-elasticity for commodities in group 2. This raises $\bar{q}_2$, moving the point into region A, and it becomes optimal to produce commodities from group 1 alone. However, both (53a) and (53b) are possible Nash equilibria, and it is therefore possible tha *the high elasticity group is produced in equilibrium when the low elasticity one shoulc have been.* If the difference in elasticities is large enough, the point moves into region C, where (53b) is no longer a Nash equilibrium. But, owing to the existence of a fixed cost, a significant difference in elasticities is necessary before entry from group 1 commodities threatens to destroy the "wrong" equilibrium. Similar remarks apply to regions B and D.

Next, begin with symmetry once again, and consider a higher c_1 or a_1. This increases $\bar{q}_1$ and moves the point into region B, making it optimal to produce the low-cost group alone while leaving both (53a) and (53b) as possible equilibria, until the difference in costs is large enough to take the point to region D. The change also moves the boundary between A and C upward, opening up a larger region G, but that is not of significance here.

If both $\bar{q}_1$ and $\bar{q}_2$ are large, each group is threatened by profitable entry from the other, and no Nash equilibrium exists, as in regions E and F. However, the criterion of constrained optimality remains as before Thus we have a case where it may be necessary to prohibit entry in order to sustain the constrained optimum.

If we combine a case where $c_1 > c_2$ (or $a_1 > a_2$) and $\beta_1 > \beta_2$, i.e., where commodities in group 2 are more elastic and have lower costs, we face a still worse possibility. For the point $(\bar{q}_1, \bar{q}_2)$ may then lie in region G, where only (53b) is a possible equilibrium and only (53a) is constrained optimum, i.e., the market can produce only a low cost, high demand elasticity group of commodities when a high cost, low demand elasticity group should have been produced.

Very roughly, the point is that although commodities in inelastic demand have the potential for earning revenues in excess of variable costs, they also have significant consumers' surpluses associated with them Thus it is not immediately obvious whether the market will be biased in favor of them or against them as compared with an optimum. Here we find the latter, and independent findings of Michael Spence in other

contexts confirm this. Similar remarks apply to differences in marginal costs.

In the interpretation of the model with heterogenous consumers and social indifference curves, inelastically demanded commodities will be the ones which are intensively desired by a few consumers. Thus we have an "economic" reason why the market will lead to a bias against opera relative to football matches, and a justification for subsidization of the former and a tax on the latter, provided the distribution of income is optimum.

Even when cross elasticities are zero, there may be an incorrect choice of commodities to be produced (relative either to an unconstrained or constrained optimum) as Figure 3 illustrates. Figure 3 illustrates a case where commodity *A* has a more elastic demand curve than commodity *B*; *A* is produced in monopolistically competitive equilibrium, while *B* is not. But clearly, it is socially desirable to produce *B*, since ignoring consumer's surplus it is just marginal. Thus, the commodities that are not produced but ought to be are those with inelastic demands. Indeed, if, as in the usual analysis of monopolistic competition, eliminating one firm shifts the demand curve for the other firms to the right (i.e., increases the demand for other firms), if the con-

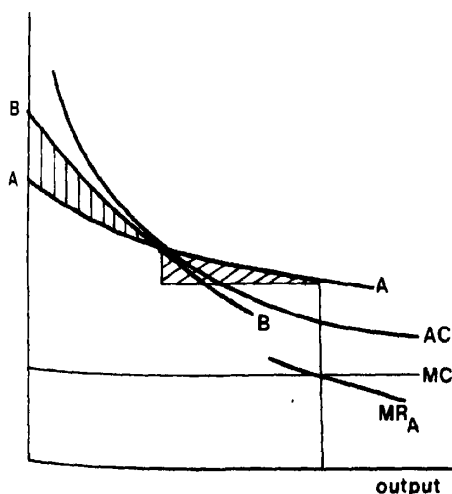



FIGURE 3

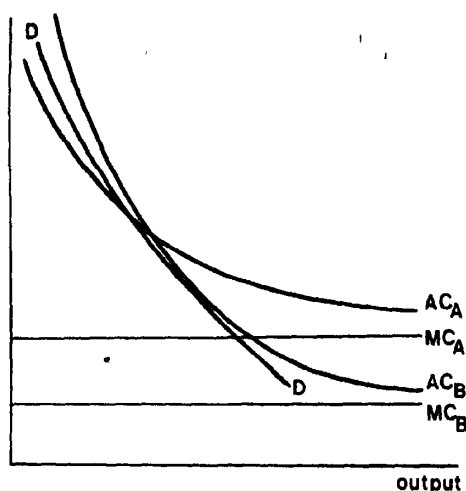



FIGURE 4

sumer surplus from *A* (at its equilibrium level of output) is less than that from *B* (i.e., the cross hatched area exceeds the striped area), then constrained Pareto optimality entails restricting the production of the commodity with the more elastic demand.

A similar analysis applies to commodities with the same demand curves but different cost structures. Commodity *A* is assumed to have the lower fixed cost but the higher marginal cost. Thus, the average cost curves cross but once, as in Figure 4. Commodity *A* is produced in monopolistically competitive equilibrium, commodity *B* is not (although it is just at the margin of being produced). But again, observe that *B* should be produced, since there is a large consumer's surplus; indeed, since were it to be produced, *B* would produce at a much higher level than *A*, there is a much larger consumer's surplus. Thus if the government were to forbid the production of *A*, *B* would be viable, and social welfare would increase.

In the comparison between constrained Pareto optimality and the monopolistically competitive equilibrium, we have observed that in the former, we replace some low fixed cost-high marginal cost commodities with high fixed cost-low marginal cost commodities, and we replace some commodities

with elastic demands with commodities with inelastic demands.

IV. Concluding Remarks

We have constructed in this paper some models to study various aspects of the relationship between market and optimal resource allocation in the presence of some nonconvexities. The following general conclusions seem worth pointing out.

The monopoly power, which is a necessary ingredient of markets with nonconvexities, is usually considered to distort resources away from the sector concerned. However, in our analysis monopoly power enables firms to pay fixed costs, and entry cannot be prevented, so the relationship between monopoly power and the direction of market distortion is no longer obvious.

In the central case of a constant elasticity utility function, the market solution was constrained Pareto optimal, regardless of the value of that elasticity (and thus the implied elasticity of the demand functions). With variable elasticities, the bias could go either way, and the direction of the bias depended not on how the elasticity of demand changed, but on how the elasticity of utility changed. We suggested that there was some presumption that the market solution would be characterized by too few firms in the monopolistically competitive sector.

With asymmetric demand and cost conditions we also observed a bias against commodities with inelastic demands and high costs.

The general principle behind these results is that a market solution considers profit at the appropriate margin, while a social optimum takes into account the consumer's surplus. However, applications of this principle come to depend on details of cost and demand functions. We hope that the cases presented here, in conjunction with other studies cited, offer some useful and new insights.

REFERENCES

R. L. Bishop, "Monopolistic Competition and Welfare Economics," in Robert Kuenne, ed., *Monopolistic Competition Theory,* New York 1967.

E. Chamberlin, "Product Heterogeneity and Public Policy," *Amer. Econ. Rev. Proc.,* May 1950, *40,* 85-92.

P. A. Diamond and D. L. McFadden, "Some Uses of the Expenditure Function In Public Finance," *J. Publ. Econ.,* Feb. 1974, *82,* 1-23.

A. K. Dixit and J. E. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," econ. res. pap. no. 64, Univ. Warwick, England 1975.

H. A. John Green, *Aggregation in Economic Analysis,* Princeton 1964.

H. Hotelling, "Stability in Competition," *Econ. J.,* Mar. 1929, *39,* 41-57.

N. Kaldor, "Market Imperfection and Excess Capacity," *Economica,* Feb. 1934, *2,* 33-50.

K. Lancaster, "Socially Optimal Product Differentiation," *Amer. Econ. Rev.,* Sept. 1975, *65,* 567-85.

A. M. Spence, "Product Selection, Fixed Costs, and Monopolistic Competition," *Rev. Econ. Stud.,* June 1976, *43,* 217-35.

D. A. Starrett, "Principles of Optimal Location in a Large Homogeneous Area," *J. Econ. Theory,* Dec. 1974, *9,* 418-48.

N. H. Stern, "The Optimal Size of Market Areas," *J. Econ. Theory,* Apr. 1972, *4,* 159-73.

J. E. Stiglitz, "Monopolistic Competition in the Capital Market," tech. rep. no. 161, IMSS, Stanford Univ., Feb. 1975.

Vertical Control By Labor Unions

By FREDERICK R. WARREN-BOULTON*

Much of the complexity of the collective argaining process can be conceptually reuced to the choice by a union of some oint along a given derived-demand curve ɔr labor. Many union actions often labeled restrictive practices" or "featherbedding" o not, however, easily fit into this frame/ork. Instead they appear to result in a /age-employment combination which lies ntirely off the derived-demand curve for ıbor. They involve either fixing the total mount of employment or fixing the laborapital (or labor-output) ratio, in addition ɒ setting the wage rate.[1] Unions have also sed more subtle measures, essentially taxes n output or capital, which can have a simiır effect to direct controls.[2]

Analytically, these actions bear a close resemblance to vertical integration, tying arrangements, or other forms of vertical control by firms with market power over a product which can be used in variable proportions as an input in a downstream production process.[3] For example, suppose that a powerful union in a competitive product industry wishes to increase the earnings per hour of its members, but is unwilling to force up the wage rate because of the output and substitution effects on employment. Vertical control can be used to reverse the substitution effect. Since formal vertical integration by unions is rare, assume the union imposes a tax or royalty on the final product. If the union wishes to increase earnings per worker with no change in employment, it can raise the royalty rate while simultaneously reducing the wage rate. The fall in the quantity of labor demanded due to the effect of the higher royalty rate on output can thus be balanced by the increase in labor demanded due to the effect of a lower wage-rental ratio on the labor-output ratio. Assuming that the elasticity of demand for the product eventually becomes greater than unity, this balancing act can continue until some finite maximum level of earnings per worker (from wages and royalties) is reached. The crucial requirement for the process to work, of course, is that the elasticity of substitution between labor and other inputs be greater than zero, since in the absence of a substitution effect there is no difference between the effects of a tax on labor and a tax on output.

This incentive for vertical control by

*Assistant professor of economics, Washington ;niversity, St Louis I would like to thank Charles ierry, Edward Kalachek, Robert Parks, and Lee Benam for helpful comments.

[1]Examples include requiring a minimum number of ıusicians per theater orchestra; minimum crew sizes nd absolute employment levels in railroading; and ·ork rules in Pacific Coast longshoring which, until ıe 1960 Mechanization and Modernization Agreeıent, required firms to employ a fixed quantity of ıbor inputs or required that labor be used in fixed roportions to other inputs or outputs. For numerous ther examples, see Lloyd Ulman, pp. 536–66.

[2]The United Mine Workers finances its Welfare and ℓetirement Fund by a royalty on each ton of coal prouced in union mines; the Mechanization and Modrnization Fund in West Coast longshoring is financed n a tonnage basis while the unemployment fund on ıe East Coast is financed by a royalty on containers roportional to the degree of anticipated labor dislacement; airline pilots are paid according to a comlex formula which closely resembles a tax on either utput or capital; and the Teamsters Union has negoıated mileage-rate differentials based on truck size and argo capacity, and a royalty payment on the transort of highway trailers on railroad flatcars. Unions nay desire such arrangements for several reasons. For xample, the airline pilot wage structure has resulted n large benefits to seniority and has facilitated barıaining in an industry with rapid productivity inreases. In addition, tying wages to particular equipnent may have enabled price discrimination by the ınion. Some union practices also regularize employment or spread the same amount of work over a larger number of employees. Nevertheless, all these measures can be used to achieve a wage-employment combination off and to the right of the derived-demand schedule for labor.

[3]For an analysis of vertical control by firms, see the author (1974).

unions has not gone unnoticed in the literature.[4] What is needed, however, is an explicit model of vertical control by unions, based on some reasonable union objective function, which can specify optimal union policy toward vertical control, analyze the determinants of such a policy, and establish the theoretical relationships between vertical control by unions and vertical control by firms. Before presenting such a model, however, we must first establish an objective function for labor unions and ask if any special constraints are applicable to the labor union case.

We assume that a union's objective function is representable by a utility function, $U = U(W, L)$ whose only components are earnings per worker and employment.[5] Graphically, the union selects that wage which results in the wage-employment combination where the derived-demand schedule for union labor is tangent to the union's highest indifference curve. In the context of a labor monopoly, vertical control can best be viewed as a method of shifting outward the demand curve for union labor. Assuming that neither earnings per worker nor employment are inferior goods to the union, the effect of vertical control on earnings per worker and on employment will lie between one of two extremes. First, the union could take all the gains from vertical control in the form of increased employment, leaving earnings per worker unchanged (a horizontal wage-preference path). At a second extreme, the union could leave employment unchanged and take all gains in the form of increased earnings per worker (vertical wage-preference path). In most cases, a relatively horizontal wage-preference path could be expected until existing union members are fully employed. Once full employment is reached, a union could be expected to use vertical control mainly to increase earnings per worker.

What factors could limit the use of vertical control by a union? The union cannot force the price of the final product above what it would cost to produce that product without the use of union labor. If the union has complete control over labor supply, this maximum price would be the cost of production when no labor is used. If a constant elasticity of substitution (*CES*) production function is assumed, the cost of production without the use of one factor is finite only if the elasticity of substitution is greater than unity. In many cases, however, the power of a union to increase production costs may also be constrained by the existence of nonunion labor. If the union forces the cost of production up too far, new firms using nonunion labor may enter the industry or existing firms may switch over to nonunion labor. The cost of production using nonunion labor will then set an upper limit for the price of the final product after vertical control.

Even if the union has effective control over labor supply, however, the imposition of vertical control could involve enforcement costs. If a strike is necessary, the union will incur a cost in foregone earnings over the contract period. These costs can be expected to be greater, the greater the loss to employers in reduced profits. Recognition of these costs results in an additional constraint on the extent of attempted vertical control.

Having specified the objective function and constraints, I proceed first to set out the simplest form of the union model, assuming a Cobb-Douglas production func-

[4] The clearest exposition is by Gregg Lewis (1951):

> The law and social canons against extortion . . in effect require labor monopolies to depend chiefly upon wage taxes for their monopoly gains. But employers can avoid wage taxes to some extent by substituting untaxed services for the taxed labor services.
>
> Unions remedy or partially remedy this defect of a wage levy in two ways. First, they supplement wage taxes with other levies, output royalties, for example, which do not have this defect. Second, they place hurdles in the way of the substitutions in order to make wage taxes approximate excise taxes in their revenue effects. These hurdles include standby rules, work demarcation regulations, and other "make-work" and "featherbedding" rules. [p. 284]

[5] For discussions of trade union goals, see Allan Cartter and Ray Marshall, pp. 240-58, and Wallace Atherton.

tion with constant returns to scale.[6] The model is then generalized to a *CES* production function, nonunion labor and enforcement costs are introduced, and the model is applied to union vertical control in coal mining.

I. Union Vertical Control With a Cobb-Douglas Production Function

Assume a Cobb-Douglas production function with constant returns to scale:

$$X = YL^{\delta}K^{(1-\delta)} \tag{1}$$

where X = final good, L = unionized labor, K = capital (or other inputs), Y = efficiency parameter, and $0 < \delta < 1$.

Assume also a constant-elasticity demand function for the final product of the form:

$$X = Z/P_x^{\eta} \tag{2}$$

where $Z > 0, \eta > 1$.

The X-industry is assumed competitive, setting price equal to marginal cost:

$$P_x = M_x = \frac{P_l^{\delta} P_k^{(1-\delta)}}{Y\delta^{\delta}(1-\delta)^{(1-\delta)}} \tag{3}$$

where M_x = marginal cost of X, P_l = wage rate, and P_k = rental price of capital.

Assume also a constant-elasticity supply function for K of the form:

$$K = H P_k^{e} \tag{4}$$

where $H > 0$, and e = elasticity of supply of K. Using equations (1)–(4) we can solve for the derived-demand curve for labor before vertical control:

$$L = C/P_l^{E} \tag{5}$$

where E, the elasticity of derived demand for labor, is given by

$$E = \frac{\eta + e + \delta e(\eta - 1)}{\eta + e - \delta(\eta - 1)}$$

and, for convenience,

$$C = Z^{\frac{(E-1)}{\delta(\eta-1)}} Y^{\frac{(E-1)}{\delta}} \delta^{E}(1-\delta)^{\frac{(E-1)e(1-\delta)}{\delta(1+e)}} H^{\frac{(E-1)(1-\delta)}{\delta(1+e)}}$$

Facing (5), the union chooses the wage-employment combination (P_l, L) where the demand curve for labor is tangent to the union's highest indifference curve.

If the union decides to exert vertical control, it can choose any one of the forms mentioned earlier. The choice would depend mainly on transactions, enforcement, and informational costs.[7] Since in the absence of such costs all forms can produce identical results, the analysis is restricted to a royalty or tax on output.

A. *Tax on Output*

Assume that the union imposes an *ad valorem* royalty or tax, t_x, on the final product. Earnings per worker will then be the

[6] The Cobb-Douglas case is presented in detail, both because it permits the derivation of explicit equations for the effects of vertical control, and because one of the most interesting cases of union vertical control occurs in the coal mining industry. The industry demonstrates strong evidence of constant returns to scale and an elasticity of substitution between capital and labor of approximately unity.

[7] Uncertainty, technical change, lack of precise technical information, and problems of specification and control lead one to expect that vertical control through a price mechanism, such as output taxes, would be considerably more efficient than direct-control measures such as wage-employment agreements or setting labor-capital ratios. Over time, direct controls can lead to a degree of inefficiency that approximates simple work-sharing. This appears to have happened in West Coast longshoring, and helps to explain the union's willingness to replace the existing work rules with more efficient forms of work-sharing, such as leisure, combined with output royalties. On the other hand, all labor earnings are in the form of wage payments if direct-control measures are used. The use of output royalties (and, to a lesser extent, taxes on capital or an equivalent profit-sharing agreement) reduces direct wage payments and may require the channeling of a significant share of earnings through a union or outside organization. If the union leadership is not particularly powerful or trusted by members, or if royalty revenue by tradition or law is assignable only to particular uses which benefit some union members more than others (such as the Welfare and Retirement Fund in the *UMW* case), then the union may choose a higher wage and lower output tax combination than that which would maximize earnings per worker for any given employment level.

new wage rate, $\bar{P}_l$, plus each worker's share of the tax receipts, $(t_x \bar{P}_x \bar{X})/\bar{L}$:

$$W = \bar{P}_l + t_x \bar{P}_x \bar{X}/\bar{L} \tag{6}$$

With $\sigma = 1$, competition in the X-industry, and constant returns to scale, wage earnings are a constant share of the after-tax value of sales of X:

$$\bar{P}_l \bar{L} = \delta(1 - t_x)\bar{P}_x \bar{X} \tag{7}$$

Substituting (7) into (6),

$$W = \bar{P}_l\left(1 + \frac{t_x}{\delta(1 - t_x)}\right) \tag{8}$$

After the tax is imposed, firms set the marginal cost of producing X equal to their net price after tax of X. Thus we must replace equation (3) with

$$\bar{M}_x = \frac{\bar{P}_l^{\delta} \bar{P}_k^{(1-\delta)}}{Y\delta^{\delta}(1 - \delta)^{(1-\delta)}} = (1 - t_x)\bar{P}_x \tag{9}$$

Using equations (1), (2), (4), and (9), we can solve for the demand for labor after vertical control:

$$\bar{L} = (1 - t_x)^{\frac{(E-1)\eta}{(\eta-1)\delta}} C/\bar{P}_l^E \tag{10}$$

To characterize the optimal tax rate for the union, we calculate the critical points of the Lagrangian W_λ of (8) subject to (10),

$$W_\lambda = \bar{P}_l\left(1 + \frac{t_x}{\delta(1 - t_x)}\right) + \lambda\left(\bar{L} - (1 - t_x)^{\frac{(E-1)\eta}{(\eta-1)\delta}} C/\bar{P}_l^E\right) \tag{11}$$

and solve for the optimal tax rate,

$$t_x = \frac{\eta + e}{\eta(1 + e)} \tag{12}$$

Substituting (12) into (8), solving for $\bar{P}_l$ and substituting into (10), we arrive at a new demand schedule for labor as a function of earnings per worker:

$$\bar{L} = \frac{C\left(1 - \frac{1}{\eta}\right)^{\frac{\eta(1+e)}{\eta + e - \delta(\eta - 1)}}}{W^E\left(1 - \frac{1}{E}\right)^E\left(1 + \frac{1}{e}\right)^{\frac{(\eta-1)e(1-\delta)}{\eta + e - \delta(\eta-1)}}} \tag{13}$$

B. *Effect of Vertical Control on Earnings and/or Employment*

We can now solve for the proportion by which the demand curve has shifted. Suppose that the union decides to take all its gains from vertical control in increased earnings per worker, leaving employment constant (vertical wage-preference path). Using equations (5) and (13), and setting $L = \bar{L}$, we can solve for the ratio of earnings per worker after vertical control to earnings per worker before vertical control:

$$R_{W\,L=\bar{L}} = \left(\frac{W}{P_l}\right)_{L=\bar{L}} = \frac{\left(1 - \frac{1}{\eta}\right)^{\frac{\eta(1+e)}{\eta + e + \delta e(\eta - 1)}}}{\left(1 - \frac{1}{E}\right)\left(1 + \frac{1}{e}\right)^{\frac{(\eta-1)e(1-\delta)}{\eta + e + \delta e(\eta - 1)}}} > 1 \tag{14}$$

where

$$\frac{\partial R_{W\,L=\bar{L}}}{\partial \delta} < 0, \quad \frac{\partial R_{W\,L=\bar{L}}}{\partial e} < 0,$$

and

$$\frac{\partial R_{W\,L=\bar{L}}}{\partial \eta} < 0$$

Thus the less elastic the supply of other i[n]puts, the less elastic the demand for the fin[al] product, and the lower the share of lab[or] costs in total costs, the greater the propo[r]tion by which earnings per worker can [be] increased. Solving for the values of $R_{W\,L=}$ as η, δ, and e reach their limit values:

$$\lim_{\delta \to 1} R_{W\,L=\bar{L}} = 1$$

$$\lim_{\delta \to 0} R_{W\,L=\bar{L}} = \infty$$

$$\lim_{e \to \infty} R_{W\,L=\bar{L}} = \frac{\left(1 - \frac{1}{\eta}\right)^{\eta/E}}{\left(1 - \frac{1}{E}\right)} > 1$$

which approaches unity as a lower limit a[s] $\eta \to \infty$ or as $\delta \to 1$, and approaches $1/\delta$ a[s] $\eta \to 1$.

$$\lim_{e \to 0} R_{W\,L=\bar{L}} = 1/\delta$$

$$\lim_{\eta \to 1} R_{W\,L=\bar{L}} = 1/\delta$$

$$\lim_{\eta \to \infty} R_{W\,L=\bar{L}} = \frac{1 + \delta e}{\delta(1 + e)\left(1 + \frac{1}{e}\right)^{\frac{e(1-\delta)}{1+\delta e}}} > 1$$

which approaches unity as $e \to \infty$, and approaches $1/\delta$ as $e \to 0$. Thus vertical control will always enable the union to increase *earnings per worker at the previous level of* employment, unless *both* the elasticity of final demand *and* the elasticity of supply of other inputs are infinite. The maximum proportion by which earnings can be increased, $1/\delta$, results as *either* the elasticity of final demand becomes unity (its assumed lower limit), *or* the elasticity of supply of other inputs approaches zero. Because vertical control enables the union to act both as a monopolist in the final-product market and as a monopsonist toward other input suppliers, vertical control can be effective even when the price of the final product cannot be raised.

Similarly, if the union chooses to take all gains in increased employment, we can again use equations (5) and (13) to solve for the maximum increase in employment with $W = \bar{P}_l$:

$$(15)\quad R_{L\,W=\bar{P}_l} = \left(\frac{\bar{L}}{L}\right)_{W=\bar{P}_l} = \frac{\left(1 - \frac{1}{\eta}\right)^{\frac{\eta(1+e)}{\eta+e-\delta(\eta-1)}}}{\left(1 - \frac{1}{E}\right)^{E}\left(1 + \frac{1}{e}\right)^{\frac{(\eta-1)e(1-\delta)}{\eta+e-\delta(\eta-1)}}} = \{R_{W\,L=\bar{L}}\}^{E} > 1$$

C. *Effect of Vertical Control on the Price of the Final Product*

The effect of vertical control on the price of the final product depends on the union's preference between increased employment and increased earnings per worker. The general expression for the price effect is given by[8]

$$(16)\quad R_{P_x} = \left(\frac{L}{\bar{L}}\right)^{\frac{\delta(1+e)}{\eta+e+\delta e(\eta-1)}}\left[\frac{1 + \frac{1}{e}}{1 - \frac{1}{\eta}}\right]^{\frac{e(1-\delta)}{\eta+e+\delta e(\eta-1)}}$$

If the union takes all gains in increased earnings per worker, $L = \bar{L}$, and

$$(17)\quad R_{P_x\,L=\bar{L}} = \left[\frac{1 + \frac{1}{e}}{1 - \frac{1}{\eta}}\right]^{\frac{e(1-\delta)}{\eta+e+\delta e(\eta-1)}} > 1$$

If the union takes all gains in increased employment, $W = \bar{P}_l$, and

$$(18)\quad R_{P_x\,W=\bar{P}_l} = \left[\frac{\left(1 - \frac{1}{E}\right)^{\delta(1+e)}\left(1 + \frac{1}{e}\right)^{e(1-\delta)}}{\left(1 - \frac{1}{\eta}\right)^{(\delta+e)}}\right]^{\frac{1}{\eta+e-\delta(\eta-1)}} \gtrless 1$$

The price increase will be greater, the more the union chooses to take its gains in the form of increased earnings rather than increased employment—a result to be expected. What is perhaps less expectable is that if the supply of K is sufficiently inelastic, a *fall* in the price of the final product is possible when vertical control is used to increase employment. To take the extreme

[8]Using equations (3) and (9)

$$\text{(a)}\quad R_{P_x} = \frac{\bar{P}_x}{P_x} = \left[\frac{\bar{P}_l}{P_l}\right]^{\delta}\left[\frac{\bar{P}_k}{P_k}\right]^{(1-\delta)}\left[\frac{1}{(1 - t_x)}\right]$$

Using equations (4) and (7):

$$\text{(b)}\quad P_k = \left[\frac{(1-\delta)P_l L}{\delta H}\right]^{1/1+e}$$

$$\text{(c)}\quad \bar{P}_k = \left[\frac{(1-\delta)\bar{P}_l \bar{L}}{\delta H}\right]^{1/1+e}$$

Substitution of P_k from (b), $\bar{P}_k$ from (c), P_l from (5), $\bar{P}_l$ from (10), and t_x from (12) into (a) gives us equation (16).

case, if $e \to 0$, equation (18) reduces to

$$(19) \qquad R_{P_x \, W=P_l} \underset{e \to 0}{=} \delta^{\frac{\delta}{\delta+\eta(1-\delta)}} < 1$$

In essence, with $e < \infty$, vertical control enables the union to effectively monopsonize nonlabor inputs. If the gain from such monopsonization is sufficiently large, some of that gain may be passed on to consumers of the final product.

D. *Effect of Vertical Control on the Price and Quantity of Capital*

In contrast to consumers of the final product, capital (or other cooperant input) suppliers always lose as a result of union vertical control. If the union takes all gains in increased earnings, the effect on the quantity of K is

$$(20) \qquad R_{K \, L=\bar{L}} = \left(\frac{\bar{K}}{K}\right)_{L=\bar{L}} = \left[\frac{1-\dfrac{1}{\eta}}{1+\dfrac{1}{e}}\right]^{\frac{e\eta}{\eta+e+\delta e(\eta-1)}} < 1$$

while if the union takes all gains in increased employment:

$$(21) \qquad R_{K \, W=P_l} = \left(\frac{\bar{K}}{K}\right)_{W=P_l} = \left[\frac{\left(1-\dfrac{1}{\eta}\right)^{\eta}}{\left(1-\dfrac{1}{E}\right)^{\delta(\eta-1)}\left(1+\dfrac{1}{e}\right)^{\delta+\eta(1-\delta)}}\right]^{\frac{e}{\eta+e-\delta(\eta-1)}} < 1$$

with corresponding effects on P_k since, from equation (4), $R_{P_k} = R_K^{1/e}$.

While capital suppliers always lose as a result of union vertical control, their losses are lessened if the union takes its gains in increased employment ($R_{K \, L=\bar{L}} > R_{K \, W=P_l}$).

The effects of optimal vertical control by a union on the labor, capital, and output markets are illustrated in Figure 1, which assumes $\eta = e = 2$, $\sigma = 1$, and $\delta = 1/4$. For ease of exposition, initial equilibrium input and output prices have been set at unity by assuming $Y = (4/3)^{.75}$, $Z = 4$, $H = 3$, and $P_l = 1$. Facing the initial demand curve for labor, LL in Figure 1A (equation (5)), the union chooses point A, resulting in $P_l = P_k = P_x = 1$. With an optimal tax ($t_x = 2/3$ from equation (12)), the labor demand curve, with earnings per worker from wages and tax revenue on the vertical axis, shifts out to $\bar{L}\bar{L}$ (equation (13)). Since, *ceteris paribus*, the output effect of a royalty would reduce the quantity of labor demanded, a reduction in the wage

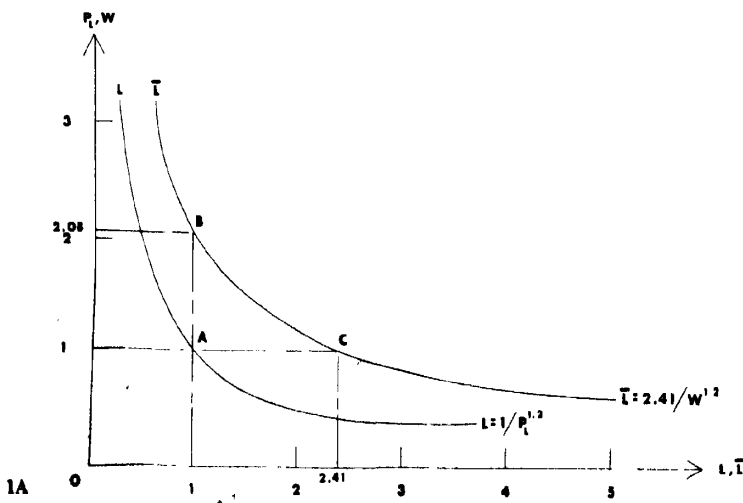
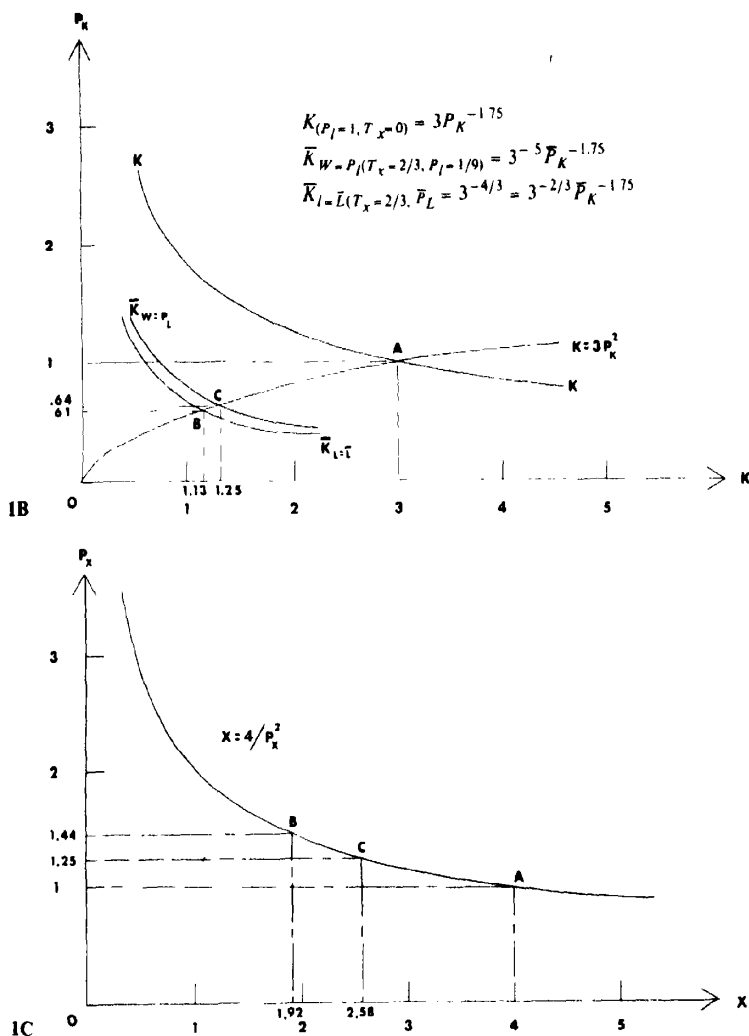



FIGURE 1: EFFECTS OF VERTICAL CONTROL BY A LABOR UNION ON THE LABOR, CAPITAL, AND OUTPUT MARKETS, ASSUMING $\eta = e = 2$, $\sigma = 1$, $\delta = 1/4$

rate is necessary if the quantity of labor demanded is to remain constant or increase after vertical control. In this example, if the wage rate is reduced to $\bar{P}_l = 3^{-4/3} \approx .23$, the output and substitution effects on the quantity of labor demanded will just balance the output effect of the royalty on the quantity of labor demanded. Since the gain in royalty revenue more than compensates for the fall in wage income, the net result is a movement to point B in Figure 1A, with a 108 percent increase in earnings per worker at the same level of employment (equation (14)). In the capital market (Figure 1B) the net effect of the royalty and the lower wage rate is a leftward shift in the demand schedule for capital,[9] resulting in a 39 percent fall

[9]Since a supply schedule for labor does not exist, we cannot solve for a true demand function for K. For any given wage rate and tax rate, however, a "demand curve" for K can be derived. Note that since we assume $\sigma = 1 < \eta = 2$ in this example, K and L are gross complements, so a reduction in the wage rate, holding the tax rate constant, shifts out the demand curve for K.

in the price of the cooperant input and a 62 percent reduction in the quantity demanded (equation (20)). In the output market (Figure 1C), the net effect of the output royalty and the reduction in both input prices is a 44 percent increase in the price of the final product.

Alternatively, if the union takes all the gains from vertical control in increased employment, with no change in earnings per worker, it can move to point C in Figure 1A by further reducing the wage rate to $\bar{P}_l = 1/9 \approx .11$. As compared with the situation before vertical control, this would result in a 140 percent increase in employment with no change in earnings per worker (equation (15)); a 36 percent reduction in the price of capital and a 68 percent fall in the quantity of capital demanded (equation (21)); and a 25 percent increase in the price of the final product (equation (18)).

E. *The Equivalence of Vertical Control by Firms and by Unions*

The introduction of a utility function for the supplier of a monopolized input eliminates both the usual profit-maximization assumption and the marginal-cost function for the monopolized input. The effects of vertical control by a utility-maximizing union can, however, be shown to be identical to the effects of downstream vertical integration by a profit-maximizing firm. Assume that good L is produced by a profit-maximizing monopolist who faces the downstream production, demand, and cooperant-input supply conditions given by equations (1) through (4). We can then solve for the effects of vertical integration on monopoly profits, the price of the final product, and the quantity of the cooperant input: all functions of g, the elasticity of the marginal-cost function for the monopolized input (see the author, 1977). As $g \to \infty$, the resulting expressions reduce to equations (15), (18), and (21), respectively. Thus if a union takes all gains from vertical control in increased employment, the percentage increase in employment is equal to the percentage increase in monopoly profits made by an integrating firm with constant marginal costs, and the effects in the final-output and cooperant-input markets are identical. As $g \to 0$, the expressions reduce to equations (14), (17), and (20). Thus vertical integration by a firm with a fixed supply of the monopolized input is equivalent to vertical control by a union with a vertical wage-preference path, where the percentage increase in monopoly profits is equal to the percentage increase in earnings per worker.

II. Generalization to a CES Production Function

The model can be further generalized by the use of a general CES production function of the form:

$$(22)\quad X = Y\left[\delta L^{\frac{\sigma-1}{\sigma}} + (1-\delta)K^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}$$
$$Y > 0, 0 < \delta < 1, 0 < \sigma < \infty$$

The correspondence between vertical control used by a union to increase earnings per worker and vertical control by a profit-maximizing firm with a fixed supply of the monopolized input still holds. The analogy between a union taking all gains in increased employment and a firm with constant marginal costs for the monopolized input, however, breaks down with $\sigma \neq 1$ Compared to vertical integration by a fir[m] with constant marginal costs, vertical co[n]trol used by a union to increase emplo[y]ment will result in a larger increase in t[he] price of the final product if $\sigma < 1$, and [a] smaller increase in the price of the fin[al] product if $\sigma > 1$.[10]

Simulation experiments for the CES ca[se] reveal an expectable pattern. The effect [of] vertical control on earnings, holding e[r...]

[10]With $\sigma = 1$, labor's share of total after-tax cos[ts] k_l, is constant and equal to δ. With constant η and the elasticity of demand for labor,

$$E = \frac{\sigma(\eta + e) + k_l e(\eta - \sigma)}{\eta + e - k_l(\eta - \sigma)}$$

is unchanged by vertical control. But if $\sigma > 1$, E fa[lls] after vertical control, while if $\sigma < 1$, E increases. Th[us] with $\sigma > 1$, the effect of a union taking some or [all] gains in increased employment would correspond [to] profit maximization with an implicit marginal-c[ost] schedule which was upward sloping, while with $\sigma <$ [1] the implicit marginal-cost curve would be downwa[rd] sloping.

loyment constant, rises from zero at $\sigma = 0$, aches a maximum at some $\sigma > 1$, and lls again to zero as σ approaches infinity. he possible gains from vertical control are ery sensitive to the values assumed for η nd δ. With low values for η and δ, and with roughly between 1.0 and 2.0, the gains om vertical control can be very large, even hen $e \to \infty$ so no monopsony gain ocurs.[11]

III. Special Constraints: Nonunion Labor and Enforcement Costs

In many cases, a union's ability to inrease labor costs may be effectively contrained by the entry of new firms using onunion labor or by a shift to nonunion abor by existing firms.

Suppose that new nonunionized firms will nter the industry if unionization raises the rice of the final product to some level $\hat{P}_x$. The demand curve for unionized labor beore vertical control is then kinked at a coresponding wage rate $\hat{P}_l$, where from equaion (3),

$$\hat{P}_l = \left[\frac{\hat{P}_x Y \delta^{\delta}(1 - \delta)^{(1-\delta)}}{P_k^{(1-\delta)}}\right]^{1/\delta} \tag{23}$$

For existing firms, $\eta \to \infty$ at $\hat{P}_x$. The elasticty of derived demand for unionized labor bove $\hat{P}_l$, however, is $E_{\eta \to \infty} = (1 + \delta e)/(1 - \delta)$ which is always greater than unity ut approaches infinity only as $e \to \infty$. Since unionized firms cannot pass on any ncrease in labor costs above $\hat{P}_l$ to conumers, any such increase is at the expense of quasi rents received by unionized capital uppliers. As P_l rises above $\hat{P}_l$, the marinal-cost schedule for X shifts upward for nionized firms, and output by unionized irms falls until P_k falls sufficiently to offset he increase in labor costs.

The kink in the demand curve for unionized labor before vertical control also results in a kink in the demand curve after vertical control. Suppose that before vertical control the union had chosen a wage greater than $\hat{P}_l$, so that $\eta \to \infty$ for unionized firms. The maximum increase in earnings per worker, holding employment constant, will then be given by the limit of equation (14) as $\eta \to \infty$. Even if the union had chosen a wage rate below $\hat{P}_l$, however, the price of the final product may reach $\hat{P}_x$ before vertical control is fully achieved. The increase in the price of X will then be less than that given by equation (17), and the increase in earnings per worker will be correspondingly lessened. The full increase in P_x and in W will be possible only if the union had originally chosen a wage rate sufficiently less than $\hat{P}_l$, so that $\hat{P}_x$ is greater than or equal to the unconstrained price of the final product after vertical control.

The nonunion labor constraint is effective over a lesser range if the union takes its gains in increased employment. As before, the constraint is completely effective over the $P_l \geq \hat{P}_l$ range where the maximum increase in employment is given by the limit of equation (15) as $\eta \to \infty$. Since

$$R_{P_x\, W=P_l} < R_{P_x\, L=L}$$

the range over which the constraint is partially effective is reduced if all gains are taken in increased employment. We would therefore expect that nonunion labor poses a lesser threat to unions using vertical control to increase employment.[12]

The union is in a worse position if existing firms would switch to nonunion labor if $P_l > \hat{P}_l$. If the union has chosen a wage of $\hat{P}_l$, vertical control cannot increase earnings per worker, and any increase in employment must come at the cost of a reduction in earnings per worker sufficient to keep the return to the cooperant input constant. The union cannot increase both employment and earnings per worker unless a higher wage rate was achievable before vertical

[11] For details of simulation experiments using *CES* production functions, see the author (1977).

[12] A clear implication of this analysis is that regulatory or statutory restrictions on entry at the product level, and/or governmental subsidies, strongly encourage vertical control by unions in these industries. Removal of such restrictions or subsidies can then be expected to reduce union resistance to, or even result in union support for, mechanization or technical change increasing labor productivity. The *U.S.* postal and maritime industries may be good examples.

control, but was not chosen due to the consequences for employment.

Unions may also restrict their use of vertical control if employer resistance would necessitate a strike with significant losses in foregone labor earnings. As mentioned earlier, one of several forms of vertical control with identical effects is a tax on the cooperant input.[13] Since the union effectively gains both monopoly power over the sale and monopsony power over the purchase of cooperant inputs, the suppliers of these inputs have a strong incentive to resist such a tax or its equivalent.[14] *Ceteris paribus*, the less elastic the supply of these inputs, the greater is both the union's potential gain and the expected enforcement costs. Thus a lower supply elasticity may encourage or discourage union vertical control, depending on the relative strength of unions and employers.

Two further points should, however, be considered. First, unless the firm facing the union is the owner of all the inputs in noninfinitely elastic supply, some or all of the losses to cooperant inputs, like losses to consumers, will not be incurred by the firm facing the union, and that firm will have a correspondingly lesser incentive to resist union vertical control. Second, limited vertical control can be used to increase utility for the union at no cost to the firm. The demand curve for labor is the locus of high points of a set of capital isoprofit curves. Note that unless the union has a fixed-proportions preference mapping, the demand curve for labor is not the contract curve.

IV. Vertical Control by the United Mine Workers of America

The clearest example of vertical control by a union has been the financing of the United Mine Workers (*UMW*) Welfare and Retirement Fund through a royalty on coal output.[15] Royalty payments, which reached $297 million in fiscal 1973–74, have totaled approximately $4 billion since 1950. Over the 1953–70 period, during which the royalty rate remained at 40¢ per ton, royalties averaged approximately 8.4 percent of the price of coal (f.o.b. mine), 10.9 percent of value-added in coal mining, and 24.1 percent of total wage costs in unionized mines.[16]

There is general agreement that the elasticity of substitution between labor and capital in bituminous coal production is approximately unity and that demand is inelastic.[17] When applied to my model, these estimates imply considerable restraint in the use of vertical control by the *UMW*. As equation (12) shows, even if $e \rightarrow \infty$, the actual royalty rate of 10.9 percent of value-added would have been optimal only if the elasticity of demand for unionized coal at the mine were approximately nine—far higher than any of the available estimates. The lower royalty rate may be due to a number of factors. One possibility is that the elasticity of demand would increase rapidly before the price of coal implied by

[13]For a proof, see the author (1977).

[14]A union unable to significantly move up its derived-demand schedule is thus unlikely to attempt major vertical control. The usual correspondence between unions which have achieved significant wage gains, such as the United Mine Workers and the Air Line Pilots Association (see Lewis, 1963), and unions which have exerted vertical control, however, may not appear when vertical control has been used mainly to maintain or increase employment as in the railroad and maritime industries. In such cases, the gains in earnings per worker from unionization may significantly underestimate the total effects of unionization.

[15]From a 5¢ per ton levy under the 1946 Krug-Lewis Agreement, the royalty rate rose rapidly to 40¢ per ton in 1952, where it remained for nearly twenty years until increased to 60¢ in 1971 and to 80¢ per ton in 1974. The 1974 three-year union contract, however has changed the financing method to a combined wage and output levy of 74¢ per ton and 90¢ per hour for Nov. 1974–Nov. 1975, 78¢ per ton and $1.40 per hour for Nov. 1975–Nov. 1976, and 82¢ per ton and $1.50 per hour for Nov. 1976–Nov. 1977.

[16]The *UMW* has also been relatively successful in raising the wage rates of its members, although at a significant cost in employment. Lewis (1970) estimated that during the 1960's, unionism raised the wage rate of union workers relative to nonunion workers by about 65–70 percent, and increased the average wage by about 32–40 percent relative to what wage rates would have been in the absence of the union. During the 1945–68 period, however, employment (man-days worked) in bituminous coal fell by about 72 percent.

[17]For estimates of the elasticity of substitution, see G. S. Maddala, Lewis (1970), and Morris Goldstein and Robert Smith. For demand elasticity estimates, see Nallapu Reddy, and Goldstein and Smith.

he "optimal" tax was reached, due to interuel competition over the upper ranges of he demand schedule. The recent large in:reases in the price of oil, however, should iave shifted any such elastic section signifi:antly upwards. A second factor may be the nstitutional constraint that royalty revenue :an be used only for the Welfare and Reirement Fund.[18] The shift in the 1974 conract to a combined output and wage-tax inancing formula, however, weakens this irgument. A third limiting force may be the hreat that a higher output tax might in'rease the percentage of bituminous coal)utput produced in nonunion mines signifi'antly above the present 25 to 35 percent evel.[19] Finally, the *UMW* simply may not)e fully aware of the advantages it has :ained or could gain from its use of an out)ut levy.

Despite its apparent underutilization, the 'ffects of vertical control by the *UMW* have)een quite significant. We begin by separatng bituminous coal into underground and urface production.[20] Following Goldstein ind Smith, we assume a constant elasticity)f demand for all coal at the delivered tage. The split between undergroundnined coal C_u, and surface-mined coal C_s, s then determined as a function of the relative delivered prices of the two types of coal $\hat{P}_s$ and $\hat{P}_u$,

$$\frac{C_u}{C_u + C_s} = \alpha(\hat{P}_s/\hat{P}_u)^\beta \tag{24}$$

We have no way of knowing the wage rate which would have been set by the union in the absence of the output royalty. We can, however, generate results for any given $P_l/\bar{P}_l$ (the ratio of what the wage rate would have been without the royalty to the actual wage rate with the royalty),[21] and thus trace out the derived-demand curve for labor in the absence of vertical control. Since data on value-added and wages are available only from the Census of Mineral Industries, we are restricted to the years 1954, 1958, 1963, and 1967.[22]

The effect of the royalty on the earnings-employment level for 1967 is shown in Figure 2, where *DD* is the derived demand curve for union labor in bituminous (underground and surface) coal in the absence of vertical control and *A* is the actual earnings-per-worker and employment combination in 1967. If the union had not utilized an

[18]This institutional constraint results in a clear conict of interest between older (especially retired) mem'ers and younger members of the union, particularly ince, until recently, the program has been financed on pay-as-you-go basis rather than being funded. A ignificant increase in the royalty rate would thus robably require that a large share of the royalty re'eipts be paid out concurrently to working members.

[19]Increasing the royalty would improve the relative osition of the smaller, narrow-seam mines with low utput-labor and capital-labor ratios, where the prob'm of nonunion labor is acute. Nonunion labor is also major problem, however, in surface mining where an icrease in the royalty would be particularly burdenome due to the high output-labor ratio and lower rice per ton.

[20]A unitary elasticity of substitution between labor nd capital can be safely assumed for both underround and surface production functions. However, ince value-added per ton of surface coal is lower than iat of underground, the union royalty rate in ad alorem terms is higher for surface than for underround coal ($t_s > t_u$). In addition, labor share in valuedded is higher for underground production than for urface ($\delta_u > \delta_s$).

[21]We assume the same value of $P_l/\bar{P}_l$ for both underground and surface mining. While this assumption reduces the estimated effects of vertical control, it appears unlikely that the *UMW* could or would deliberately adjust relative wage rates in surface and underground mines so as to maximize the gains from vertical control.

[22]Since a description of the methods used to calculate the parameter values from Census data and the derivation of the equations used would be rather lengthy, they are not reported here but are given in my forthcoming study (1977). Two assumptions, however, should be noted First, I have assumed an infinite elasticity of supply for all cooperant (nonlabor) inputs. To the extent that this elasticity is less than infinite, the quantitative results reported below are an underestimate of the effects of vertical control by the *UMW*. On the other hand the only adjustment for the existence of nonunion mines has been to divide the final demand elasticities for underground and surface coal by the fractions of coal produced in unionized underground and surface mines. This assumes that vertical control by the union has not increased the percentage of coal produced in nonunion mines, an assumption which can be expected to result in an overestimate of the effects of vertical control. In the absence of accurate information as to either the elasticity of supply of cooperant inputs or the effect of unionization on nonunion shares, it is at least comforting that the two biases act in opposite directions.

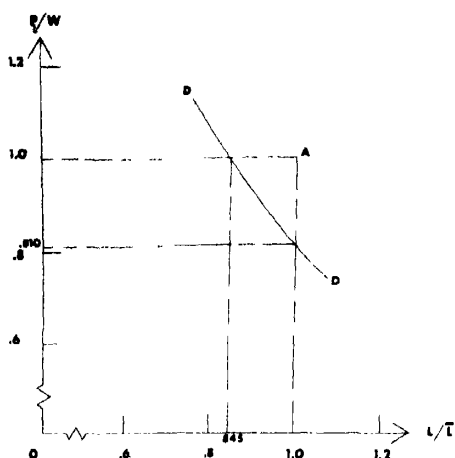



FIGURE 2· EARNINGS-EMPLOYMENT EFFECT OF THE *UMW* ROYALTY, 1967

TABLE 1—VERTICAL CONTROL USED TO INCREASE EARNINGS PER WORKER ($L = \bar{L}$)

	1954	1958	1963	1967
P_l/W	0.853	0.853	0.819	0.810
$P_l/\bar{P}_l$	1.027	1.021	1.015	1.020
$L_u/\bar{L}_u$	0.987	0.986	0.981	0.982
$L_s/\bar{L}_s$	1.136	1.104	1.111	1.117
$L/\bar{L}$	1.000	1.000	1.000	1.000
$K_u/\bar{K}_u$	1.013	1.007	0.996	1.001
$K_s/\bar{K}_s$	1.166	1.128	1.128	1.139
$C_u/\bar{C}_u$	0.996	0.995	0.989	0.992
$C_s/\bar{C}_s$	1.154	1.119	1.123	1.132
$C/\bar{C}$	1.032	1.029	1.036	1.036
$P_u/\bar{P}_u$	0.913	0.926	0.921	0.917
$P_s/\bar{P}_s$	0.854	0.875	0.861	0.864
$\hat{P}_u/\bar{\hat{P}}_u$	0.953	0.959	0.955	0.952
$\hat{P}_s/\bar{\hat{P}}_s$	0.937	0.945	0.936	0.934

output royalty, achieving the actual 1967 level of employment would have required a 19.0 percent cut in earnings per worker. Alternatively, achievement of the actual 1967 level of earnings per worker without the royalty would have resulted in unemployment for 15.5 percent of the 1967 unionized work force.

The effects of vertical control on underground and surface coal production levels, prices, and capital levels depend on the particular point along *DD* which the union would have chosen if the Welfare and Retirement Fund had been financed by a wage levy rather than an output royalty. Table 1 shows what would have happened if the union had attempted to achieve the actual employment levels in each year without using a royalty. In 1967, for example, earnings per worker would have been 19 percent lower (the net effect of the loss in royalty earnings and an increase of 2 percent in wage rates). While total unionized employment in underground and surface mines would be unchanged, underground-mine employment would have been 1.8 percent lower and surface-mine employment 11.7 percent higher. The capital stock in underground mines would have been only 0.1 percent higher, but capital in surface mining would have been 13.9 percent greater than the actual stock in 1967. Underground coal production would have been reduced by 0. percent, while surface production woul have been 13.2 percent higher. Total ton nage (underground and surface, not ad justed for differential *BTU* content) woul have been 3.6 percent higher. Value-adde prices would have been 8.3 percent lowe for underground coal and 13.6 percen lower for surface coal. Delivered price would have been 4.8 percent lower for underground coal and 6.6 percent lower for surface coal.

Table 2 shows the corresponding effects if the union had attempted to achieve the actual earnings-per-worker levels in each year without using a royalty. It is particularly interesting to note that the resulting fall in employment would have occurred entirely in underground mining, with surface-mining employment actually higher for all years. Comparing Tables 1 and 2 also shows that use of an output royalty results in a considerably larger reduction in capital utilization if vertical control is used to increase employment.[23]

[23]The contrast with the conclusion of Section ID *supra* is worth noting. A finite "optimal" output levy requires that $\eta > 1$. Since we assume $\sigma = 1$, capital and labor will be gross complements if $\eta > 1$, so that the Section ID result holds. If $\eta < 1$, however, as in the *UMW* case with a "suboptimal" royalty rate, capital and labor will be gross substitutes, and vertical control will result in a larger reduction in the capital stock *if the royalty is used to increase employment.*

TABLE 2—VERTICAL CONTROL USED TO INCREASE EMPLOYMENT ($W = P_l$)

	1954	1958	1963	1967
P_l/W	1.000	1.000	1.000	1.000
$P_l/\bar{P}_l$	1.203	1.197	1.239	1.259
$L_u/\bar{L}_u$	0.865	0.864	0.825	0.819
$L_s/\bar{L}_s$	1.152	1.067	1.017	1.017
$L/\bar{L}$	0.891	0.888	0.852	0.845
$K_u/\bar{K}_u$	1.041	1.034	1.022	1.031
$K_s/\bar{K}_s$	1.386	1.277	1.261	1.281
$C_u/\bar{C}_u$	0.924	0.936	0.919	0.926
$C_s/\bar{C}_s$	1.291	1.196	1.179	1.195
$C/\bar{C}$	1.007	1.007	1.010	1.009
$P_u/\bar{P}_u$	1.012	1.011	1.017	1.012
$P_s/\bar{P}_s$	0.907	0.927	0.916	0.921
$\hat{P}_u/\bar{\hat{P}}_u$	1.006	1.006	1.009	1.007
$\hat{P}_s/\bar{\hat{P}}_s$	0.961	0.969	0.962	0.963

One intriguing possibility which emerges from these results is that underground coal mine operators may actually have benefited from the *UMW*'s use of an output royalty. If vertical control has been used to increase employment, Table 2 shows that the reduction in the wage rate more than offsets the effect of the output royalty on underground unit costs. Combined with the outward shift of the demand curve for underground coal (due to the increase in the relative price of surface-mined coal), this reduction in underground unit costs results in an approximately 7 percent increase in underground production over the period. If vertical control has been used to increase earnings per worker (Table 1), unit costs for underground mines have risen, but the demand shift is still sufficient to result in a marginal increase in underground output. The implication is that as long as the same wage rate and royalty per ton is imposed on all coal mines, underground coal operators as a group (and especially the operators of small underground mines with high labor/output and labor/capital ratios) could well have benefited from the *UMW*'s use of an output royalty.[24]

V. Conclusions

Restrictive work rules, output taxes, or other forms of vertical control can be of significant benefit to union members, particularly when the elasticity of demand for the final product, the share of labor in total costs (more precisely, the distribution parameter), and the elasticity of supply of other inputs are all low, while the elasticity of substitution between labor and other inputs is somewhat greater than unity. The effects of such union actions are analytically very similar to the effects of vertical integration, tying arrangements, or other forms of vertical control by firms with market power over a product which can be used in variable proportions as an input by downstream producers. Since estimates of substitution elasticities between capital and labor are usually higher than between commodity and other inputs, however, we would expect vertical control to be more attractive to unions than to firms.

Despite the potential gains, unions may avoid or limit their use of vertical control if the parameter values for η, δ, e, and σ are such that vertical control would be of minor benefit, if nonunion labor poses a major problem, or if strong resistance by employers would require significant enforcement costs for the union. Finally, many unions may simply be unaware of the potential gains from vertical control.

[24] Williamson (1968) has argued that higher wage rates imposed on all mines may be in the interest of the operators of the larger underground coal mines where labor-capital ratios are significantly lower than in the smaller underground mines. The introduction of surface mines as a third sector with the lowest labor-capital ratios would weaken and perhaps invalidate this conclusion. A comparison of Tables 1 and 2, however, indicates that Williamson's proposition is correct for surface-mine operators.

REFERENCES

Wallace N. Atherton, *Theory of Union Bargaining Goals*, Princeton 1973.

Allan M. Cartter and F. Ray Marshall, *Labor Economics: Wages, Employment and Trade Unionism*, Homewood 1972.

M. Goldstein and R. S. Smith, "The Predicted Impact of the Black Lung Benefits Program on the Coal Industry," in Orley Ashenfelter, ed., *Evaluating the Labor Market Effects of Social Programs*, Princeton forthcoming.

Harold G. Lewis, "The Labor Monopoly Problem: A Positive Program," *J. Polit. Econ.*, Aug. 1951, *59*, 277–87.

———, *Unionism and Relative Wages in the United States: An Empirical Enquiry*, Chicago 1963.

———, "Unionism, Wages and Employment in U.S. Coal Mining, 1945–68," abstracted in *West. Econ. J.*, Sept. 1970, *8*, 318.

G. S. Maddala, "Productivity and Technological Change in the Bituminous Coal Industry, 1919–1954," *J. Polit. Econ.*, Aug. 1965, *73*, 352–65.

N. N. Reddy, "The Demand for Coal: A Cross-Sectional Analysis of Multi-Fuel Steam Electric Plants," *Ind. Org. Rev.*, 1975, *3*, 37–42.

Lloyd Ulman, *The Rise of the National Trade Union: The Development and Significance of Its Structure, Governing Institutions, and Economic Policies*, Cambridge 1966.

Frederick R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, July/Aug. 1974, *82*, 783–802.

———, *Vertical Control of Markets: Business and Labor Practices*, Cambridge forthcoming 1977.

O. E. Williamson, "Wage Rates as a Barrier to Entry: The Pennington Case in Perspective," *Quart. J. Econ.*, Feb. 1968, *82*, 85–116.

Black-White Male Wage Ratios: 1960–70

By JAMES P. SMITH AND FINIS R. WELCH*

It is likely that at no time in recorded history has racial equality received as much attention as during the 1960's in the United States. It is also likely that economists will associate the last decade most closely with the beginning of a dramatic erosion in some of the more historically persistent wage differentials in the United States. This paper documents changes in black-white male earnings[1] as these changes are revealed in two surveys, the 1 in 100 Public Use Samples of the Census of Population in 1960 and 1970. However one slices the data, the improvement in the relative income of black males during the 1960's is impressive. Equally eye-catching is the universal sharing of these gains across experience and schooling classes. But even though all black age and experience groups gained on average relative to whites, the gains were not equally distributed: those whose relative position improved most are more likely to be the most educated and the more recent entrants into the labor market.

There are three sections to this paper. The first summarizes relative black-white earnings and wage ratios by schooling class and estimated time out of school (work experience) for both 1960 and 1970. The second section focuses on two issues that have occupied much of the recent research on black-white incomes: 1) the viability of schools as a mechanism for increasing black economic status; and 2) the effectiveness of governmental antidiscrimination legislation in eliminating racial wage differentials. Recent work has challenged the earlier pessimism that formal schooling was not an effective means of improving the well-being of blacks. In addition to our assessment of this controversy, we will evaluate a number of hypotheses related to the schooling question, i.e., secular and regional variation in school quality, and life cycle components to the return to schooling. Many[2] have argued that the rise in black-white ratios over the 1960's is the effect of enforced compliance to fair employment legislation popularly known as "affirmative action." While this issue is not addressed directly, we have compiled some indirect evidence that the economic impact of this legislation is probably overrated. Based on our work, the largest gains in black-white wage ratios have occurred in industries least vulnerable to federal or local government "arm-twisting." The final section presents a partial accounting giving order-of-magnitude estimates of some sources of black-white income differences as of 1970 and of changes between 1960 and 1970.

I. Income and Wage Ratios: An Overview

In this section we highlight those patterns of black-white earnings ratios that capture the major changes of interest during the 1960's. The ratios of black-white wages are listed in Table 1 in a manner designed to separate the distinct life cycle and cohort

*Economist, The Rand Corporation; professor of economics, University of California at Los Angeles and economist, The Rand Corporation, respectively. This research was supported by a contract from the U.S. Department of Labor and a grant from HEW to The Rand Corporation, and by a grant from the Ford Foundation to the National Bureau of Economic Research. Frank Berger, Richard Buddin, Ann Dukes, Bill Gould, and Iva MacLennan provided competent research assistance.

[1]The samples used here refer to males with earnings in 1959 and 1969. Self-employed are excluded, as are persons whose imputed years of work experience exceeds 40 years or is negative. Experience is defined as current age minus age at leaving school. Following Giora Hanoch, the following school-leaving ages were assumed.

	Schooling Level						
	0-7	8	9–11	12	13–15	16	17+
Age.	14	16	18	20	23	25	28

[2]See especially Richard Freeman and Wayne Vroman. For a critical examination of the Freeman paper, see Robert Flanagan.

TABLE 1—BLACK-WHITE EARNINGS RATIOS FOR COHORTS IN 1960 AND 1970

Cohort Experience as of 1970 (years out of school)	Average Annual Earnings 1970	1960	Average Weekly Earnings 1970	1960	Within[a] Cohort 1970–60	Between[b] Cohort 1970–60
I. All School Completion levels						
1 5	.653	-	.702	—	—	.134
6 10	.648	—	.677	—	—	.104
11 15	.621	510	.641	.568	073	.060
16 20	.601	.529	.618	573	.045	.031
21 30	594	.545	.616	.585	.031	.042
31 40	.604	.540	.620	.574	.046	.046
II. Elementary School Graduates (8 years completed)						
1 5	835		.865	—	-	162
6 10	779	—	802	—	—	089
11 15	.708	.673	737	703	.034	013
16 20	710	.688	.717	.713	004	.021
21 30	749	.671	763	708	055	.022
31 40	721	719	740	741	- 001	030
III High School Graduates (12 years completed)						
1 5	.775	—	806	—	-	.092
6 10	.769		.791	-	—	077
11 15	729	.654	749	714	035	.067
16 20	731	676	750	714	036	060
21 30	.678	655	.698	685	013	050
31 40	.675	623	690	648	042	100
IV College Graduates (16 years completed)						
1 5	716	-	775	-		.120
6 10	647	—	692	-	-	110
11 15	.662	618	.688	655	033	106
16 20	654	559	675	582	093	.158
21 30	519	446	557	470	087	.136
31 40	504	389	.522	.421	101	100

[a]First difference of the corresponding entries in columns 3 and 4

[b]First difference of entries in columns 3 and 4 of groups with same experience, e g 134 = 702 – .568.

trends.[3] The 1970 column gives black-white ratios for the six experience classes in 1970. The second column gives the same ratios for 1960, but it is pushed down by two rows. Thus, the first entry .510 is black-white earnings ratio for the 1 5 experience group in 1960; this cell had 11-15 years of experience in 1970. The trend within a cell as a new cohort enters can be read up the diagonal. The within-cohort life cycle trends are illustrated across a row. A number of patterns are apparent. First, the large earnings differentials that existed in 1960 were partly eroded between 1960 and 1970, but as of 1970 differences remained large Second, black-white earnings ratios are highest for those who entered the labor market during the 1960's, and they are higher for those entering between 1965 and 1970 than for 1960-65 entrants. Among cohorts who were in the labor market in 1960, with the exception of college graduates, we find that by 1970 the relative position of blacks had improved only slightly over 1960 levels. But, among the cohorts whose work experience predates 1960, the pattern exhibited for post-1960 entrants continues to hold: namely, that younger black cohorts—more recent entrants into the labor market—fare better than their earlier counter-

[3]Numbers reported are ratios of averages, i.e., they are average black earnings or weekly wages relative to appropriate averages for whites Weekly wages are earnings last year divided by weeks worked last year. The average weekly wage used here is total earnings of all persons divided by total weeks worked, i.e., individual earnings per week are weighted by weeks worked.

parts. Third, the gains that occurred between 1960 and 1970 are broadly based. Wage growth was fairly uniformly distributed across experience and education cells for white males, but this apparent growth neutrality for whites contrasts sharply with the trends emerging among blacks where the extent of the gain is positively related to education level and to time of entry into the labor market.[4] The most spectacular improvement is undoubtedly that of college-educated blacks.

The observed patterns in black-white earnings ratios of persistent cohort improvement, relatively larger gains to more schooled workers, and a smaller rate of increase within cohorts is not consistent with either a pure life cycle or vintage hypothesis. The frequently observed cross-sectional decline in black-white wage ratios as age is increased produced theories of labor market discrimination that favored a life cycle explanation. The presumption underlying such theories is that some jobs, those dubbed "secondary," are dead-end, with little prospect for career progress in wages and job status, while other jobs facilitated upward mobility. Persons who seemed likely candidates for secondary careers were disproportionately black and less schooled. In either cross section (reading down a column in Table 1), black-white wage ratios clearly deteriorate as experience increases or vintage decreases and the rate of decline is more pronounced at higher levels of school completion. But, the within-cohort changes between 1960 and 1970 are the mirror image of the cross section. Not only did the relative position of blacks improve as they added 10 years of work experience, but this improvement was greatest at higher schooling levels. This data rejects the "secondary labor market" view and lends more support to the alternative that differences in the cross section are indicative of cohort differences.

We attribute the rising wages between cohorts to differential vintage effects that favor black males. The simplest vintage model would describe black-white wage ratios as functions only of cohort— of time of entry into the job market.[5] If vintage effects reflect secular change either through rising relative quality of black labor or declining front-end labor market discrimination, younger, more recent cohorts of blacks would fare better in comparison to whites than older cohorts, but the differences existing within a given cohort in 1960 would persist until 1970. Our "best guess" for rationalizing the pro-skill bias in rising black-white wage ratios within cohorts is increasing school quality. There is evidence that nominal attributes such as days attended, school retardation rates, teacher educational levels, and teacher salaries have been improving throughout most of this century for black students relative to whites (see Welch, 1973a, b). More importantly, black students have been switching into integrated traditionally white-dominated schools—especially colleges.

Calendar year effects, i.e., changes in labor markets, emerge as a likely candidate for explaining the observed increases within cohorts. An obvious source of gain between 1960 and 1970 is the improvement in the general level of economic activity.[6] Indeed, it is important to establish that the black achievements were not solely due to business cycles. The penalties imposed by business contractions are not uniform over

[4]To illustrate the separate white and black trends, consider the following table which gives by race the ratio of 1970 to 1960 weekly wages (in constant dollars) by separate experience and education groups.

Experience class:		1-5	11-15	21-30
Elementary	White	1.23	1.13	1.21
	Black	1.51	1.14	1.17
High School	White	1.26	1.20	1.21
	Black	1.46	1.32	1.29
College	White	1.27	1.23	1.23
	Black	1.61	1.55	1.63
Complete Sample:				
	White	1.30	1.30	1.32
	Black	1.61	1.45	1.40

[5]For example, in the well-known human capital model proposed by Yoram Ben-Porath, the life cycle wage growth paths are independent of the initial human capital stock.

[6]The *U.S.* aggregate average unemployment rate was 5.5 percent in 1959 and had fallen to 3.5 percent by 1969. We have also excluded males with zero earnings from this study.

education, age, or racial groups, and as business conditions improved over the decade, black earnings would have increased relative to whites. Other researchers have provided convincing evidence that during recessionary periods those most adversely affected are the less skilled (schooled) and, symmetrically, these same people gain most during business expansions. Yet, in our wage comparisons, those blacks who gained most in comparison to whites had the most schooling. Secondly, the change in the real characteristics of people (i.e., schooling or location) that we observed during the decade would, in the absence of any business cycle trends, have led to an increase in the relative income position of blacks. Third, comparisons of occupational "sameness" between blacks and whites over the decade indicate that the observed wage changes are likely to be permanent. Suppose, for example, that the income gains of blacks had been achieved alongside increasing black-white occupational disparity (segregation). In contrast to a situation in which black and white occupational and industrial distributions were increasingly congruent, we would be quite reluctant to argue that we are moving toward racial parity. According to our measure of occupational congruency, as of 1970, the occupational distribution of black males who entered the work force between 1960 and 1970 is more similar to the white distribution than to the 1970 occupational distribution of blacks who had entered the work force prior to 1960.[7] Because of changes in economic conditions between 1960 and 1970, the agreement between observed wage changes and the presumably more cyclically durable occupational distributions is taken as support for the notion that wage comparisons reflect, at least in part, long-run changes.,

A. *The Statistical Frame*

Because our overall objective is to account for differences in black and white earnings as of 1970 and for changes in earnings ratios between 1960 and 1970, we have organized the data around an internally consistent set of ordinary least square (*OLS*) regressions. Regressions are estimated separately for six work experience classes, but data for a given class are pooled among years (1960 and 1970) and across races (black and white). The regression format is:

$$(1)\quad y = x'b_0 + d_1x'\delta_1 + d_2x'\delta_2 + d_1d_2x'\delta_{12} + u$$

where $d_1 = 1$ if black, 0 otherwise; $d_2 = 1$ if 1960, 0 otherwise

The dependent variable y is the logarithm of the weekly wage in constant Consumer Price Index (*CPI*) dollars. The vector of explanatory variables x is partitioned into four groups: 1) school completion; 2) geographic location; 3) government employment; 4) years of work experience. Exact definitions of each of the included variables are provided in the appropriate sections. In this form, b_0 is the parameter vector associated with x for whites in 1970, δ_1 is the black-white difference in parameters in 1970, δ_2 is the 1960–70 change in parameters for whites, and δ_{12} is the 1960–70 change in the black-white difference in parameters. In this completely interactive form the parameter estimates for each of the race-year groups are identical to those that would have been obtained from separate regressions performed within each group.[8]

Estimates of this fully interactive model were too general in that it allowed for parameter differences that apparently did

[7] Our index of congruency between the ith and jth group is defined as

$$I_{ij} = \sum_l f_{il} f_{jl} + \frac{(f_{il}^2 + f_{jl}^2)}{2}$$

where l indexes the occupation and f refers to proportions such that $\sum_l f_{il} = f_{jl} = 1$. This index is bounded by 0 (no overlap in distributions) and 1 (complete congruency). This index for pre-1960 black entrants and all whites is 0.722. It is 0.875 for pre- and post-1960 black entrants and rises to 0.901 for the comparison of post-1960 blacks to whites.

[8] It does give slightly different test statistics since in this pooled form the estimate of residual variance (σ_u^2) is based on the sum of the residual quadratics over the four groups instead of being estimated separately for each group.

not exist. We found in our initial attempts at accounting that too often "statistically significant" effects were numerically swamped by estimates that contained too much error to be believed. We therefore imposed a number of exclusion restrictions on parameters. In all cases, these restrictions are for "interaction" effects and the procedure followed was generally that of deleting "statistically insignificant" effects.[9] Such a procedure risks incorrect inference because sequential tests are not independent. We note only that the constrained estimates are more efficient and whatever biases they imply are necessary to clarify our estimates of factors contributing to increased black-white wage ratios.

II. Schooling and Affirmative Action

We focus initially on two "problems" that have been a central part of most studies of racial wage differences—the income returns to schooling and the impact of governmental antidiscrimination legislation. There are three important questions regarding the impact of schooling on earnings: 1) are returns for blacks as high as for whites; 2) can inferences of secular change in quality of schooling be drawn by observing coefficient differences by race, among work experience classes, and through time; and 3) does the post-1970 evidence of deteriorating returns to higher education support the notion that whatever its historic role, education is unlikely to be a viable route to economic mobility in the future. While regression of *log* earnings on schooling can never definitively resolve questions concerning the underlying cause of the income schooling relationship, we believe that any observed regularities in the pattern of schooling coefficients across age, race, or cohort groups are at the least suggestive about the nature of the schooling-income relationship. For affirmative action, we examine the apparent contributions of governments to black wage growth by contrasting performance between direct government employment, employment in regulated industries, and indirect government employment (work in private industries whose product is purchased by governments).

A. *Schooling*

The estimated equations include two variables for school completion. The first ranges from 0 to 12 and indicates years of elementary and secondary schooling. The second measures years of post-secondary schooling. If a person reports a positive number of years of college, the grade school variable is set equal to 12.[10] This "spline" function is linearly segmented to permit slope coefficients to differ between the first 12 and succeeding years, but the linear segments are constrained to join at 12 years. Tests of equality for the two coefficients within experience classes show that it can be rejected in most cases.

In examining the unconstrained estimates for the fully interactive model, we found no significant evidence that returns to elementary and secondary schooling differed between 1960 and 1970. In each year, grade schooling coefficients were higher for whites than for blacks. In contrast, returns to college were higher in 1970 than in 1960 and

[9]We do not report here either the full set of constrained or unconstrained estimates. These are available in Smith-Welch, p. 59 Several of the variables suppressed in the constrained estimates are statistically significant in the fully interactive model. Although the imposed constraints delete variables that in the main appear insignificant in the fully interactive model, the joint test for significance clearly rejects the null hypothesis The computed F-statistics are:

Experience Class	F-Statistic	Degrees of Freedom
1-5	1.95	28; 17,613
5 10	1.92	29; 17,413
11-15	1.18	20; 16,722
16-20	1.36	33; 16,576
21-30	1.51	33; 31,254
31-40	2.61	33; 24,899

The associated (0.01) critical value $F(30;\infty) = 1.69$. This problem of an inability to reject hypotheses is common to large samples and has resulted in a number of attempts to weaken test criteria. See, in particular, Carlos Toro-Vizcarrondo and T. Dudley Wallace, and Wallace.

[10]This is similar in spirit to the specification proposed by Lester Thurow. He included continuous variables: 0-8, 9-12, and 13+ years of schooling. However, he then takes the *log* of these variables and does not permit interaction between experience and years of schooling.

TABLE 2—ESTIMATED SCHOOLING COEFFICIENTS[a]

Years of Experience	White		Black	
Elementary and Secondary:				
1-5	.143		097	
6-10	.100		.083	
11-15	.069		.054	
16-20	.062		.046	
21-30	.059		.032	
31-40	.050		.026	
	1970	1960	1970	1960
College:				
1-5	.124	.099	.158	.132
6-10	.093	.082	b	b
11-15	.092	.075	b	b
16-20	.088	.065	b	b
21-30	.077	070	b	b
31-40	.074	.050	b	b

[a]The schooling coefficients can be read as the fractional increase in income associated with an extra year of schooling. Since standard errors of estimation are small relative to coefficient estimates they are not reported.

[b]Constrained to equal white coefficients. Actually, in the unconstrained estimates, four of the five estimated coefficients are higher for blacks than whites in 1970, whereas the opposite is true for 1960: four of five coefficients are higher for whites. Nonetheless, numerical differences are small and are statistically insignificant.

black-white differences were slight. Table 2 reports the constrained estimates that suppress year interaction for grade schooling.

Consider first the question of black-white differences. Initial studies based on the 1960 Census by Hanoch, Thurow, and others painted a consistent picture of low returns to schooling for blacks. Later work based principally on the Survey of Economic Opportunity (see Leonard Weiss and Jeffery Williamson, 1972, or Welch, 1973a) estimated black education coefficients that were as large or even larger than those obtained for whites in 1960. This raised the possibility of a strong structural shift during the decade in the effectiveness of black schooling. Continuing this controversy in a recent issue of this *Review*, Charles Link, using group data from the 1970 Census, obtained results quite similar to those reported in Thurow's 1960 Census study. In a detailed rebuttal, Weiss-Williamson (1975) found considerably higher black returns in 1970 using individual data. Because we have pooled estimates based on both the 1960 and 1970 Census, we are able to directly test for black-white differences in the return to schooling and the existence of structura shifts. For the grade school variables, th full interactive estimates suggest that the re turns to grade school for blacks are lowe than for whites. In contrast, the margina returns to post-secondary schooling ar actually higher for blacks than for whites i the 1-5 experience interval. We find no sta tistically significant difference by race in th college returns in the other experience in tervals. If school systems are not an effec tive means of increasing black incomes, it i clear that the problem lies at the elementary and secondary levels.

Concentrating on the coefficients that measure secular movements, there was no trend in the returns to grade school for either race between 1960 and 1970, but higher wage returns to college in 1970 in all classes.[11] There is some evidence that the college coefficient rose more during the de-

[11]In a more detailed specification in which the 0-12 segment was divided into an elementary (0-8) and high school range (9-12), there seems to be support for a decline in the returns in the 0-8 range and a slight rise in the 9-12 range for both races.

cade for blacks than for whites.[12] These estimates show that earlier studies based on the 1960 Census which did not allow interactions with work experience seriously underestimated the prospects for black schooling. Although returns to college rise over the decade, there is no indication of a pronounced structural shift.

Weiss and Williamson (1972) describe a number of hypotheses relating to regional variation in the quality of black schools. As they point out, one possible explanation for a weak effect of education on black income is that historically blacks were educated primarily in low quality southern schools. If this were the cause of low payoffs to black education, one might expect black returns to increase relative to those of whites as more blacks were educated in the North. Using the 1967 Survey of Economic Opportunity (*SEO*), they include as explanatory variables in an earnings equation, an index of an individual's residence at age 16 as a proxy for the region where schooling occurred.[13] They find that blacks who lived in the North and large southern Standard Metropolitan Statistical Areas (*SMSA*) when they were 16 actually had slightly lower earnings, giving little support for regional variation in black schooling quality. They also report that the largest negative differential for those blacks living in the North at age 16 existed among younger cohorts. On this basis, Weiss-Williamson suggest that there may have been some decline in the quality of northern ghetto schools. We think these tests are deficient for two reasons: because their hypotheses relate to quality of schools, the appropriate specification involves an interaction with education rather than a dummy variable for residence; there is the problem in separating schooling quality from the market for skilled labor.[14] Since skilled labor is relatively scarce in the South, the premium to skill (education) may be higher there even though schools are of lower quality. To separate these effects, we have interacted years of education separately with dummy variables for southern residence and southern birth—the former to capture regional differences in the returns to skill and the latter to measure regional differences in quality of schooling.[15] (See Table 3.)

TABLE 3—REGIONAL RETURNS TO EDUCATION

Years of Schooling	Education	Southern Residence Educational Interaction	Southern Birth Education Interaction
1970 Whites			
0-12	.0748	.0251	-.0095
13+	.1068	.0349	-.0253
1970 Blacks			
0-12	.0604	.0140	-.0048[a]
13+	.1202	.0284	-0139[a]

[a]Indicates not significant at 95 percent confidence interval.

According to our estimates, the South is characterized by a larger return to skill for both races. The fact that schooling demands a premium in the South may well be the basis of the migration of skilled labor observed in the post-World War II era. Southern birth (presumably, attendance at southern schools) lowers the return to schooling, though not significantly so for blacks. If there is evidence of quality of schooling here, it is that southern schools are inferior for whites. The case for regional comparisons in quality of schooling for blacks is less clear.[16]

[12]In a dummy variable specification, we find that the difference between the change in the returns between 12 and 16 years of schooling for blacks and whites was for years of experience):

Years:	1-5	6-10	11-15	16-20	21-30	31-40
	.134	.096	.021	.358	.213	.160

[13]In particular, they allowed for five regional labor markets; rural South, small city South, medium city South, large city South, and non-South. We have substituted some regional detail in the South for a more precise specification of the education region interaction.

[14]See Welch's 1967 paper, the first systematic attempt to separate the effects of schooling quality and factor ratios on existing skill differentials.

[15]These regressions are computed only for comparison with the Weiss-Williamson estimates. Since all experience classes are pooled, regressions described do not conform to the experience specific estimates discussed in other sections.

[16]Thus, our conclusion is in accord with Weiss-Williamson even though our test is quite different.

Returning to patterns among experience groups exhibited in Table 2, schooling is a less important discriminator of earnings for more experienced or older vintaged workers. By looking at changes within cohorts in Table 2, we can follow a single cohort and observe the returns to schooling.[17] For the college segment of the spline, with the exception of the initial 1960 cohort, there is a slight rise in the return to schooling within cohort. In contrast, the pattern emerging within cohorts for the elementary and secondary segment is precisely the decline predicted by the cross-sectional relation. In an earlier paper, Welch (1973) performed a similar comparison using the 1960 Census and 1967 *SEO*. There, although within each cross section the education coefficients declined with experience, the within-cohort coefficients were remarkably stable. This was used as evidence that the cross-sectional pattern was mainly one of improving secular change in the quality of schools and not a life cycle phenomenon. Obviously—at least for the elementary coefficients—our experiment with the two Censuses tells a different story. One possibility is that a life cycle component was dismissed too readily. It is conceivable that the skills acquired in college are more complementary with job experience than are skills acquired in grade school. Moreover, the relatively tighter labor markets in 1969 compared to 1959 may have depressed returns to schooling and nullified long-run tendencies.

A final possibility is that there may not be any strong a priori reason to predict that increased quality of schooling will necessarily alter the semilogarithmic coefficient of wages on schooling. To illustrate, let aggregate output be

$$Y = f(g(N, QS), K) \tag{2}$$

where N is the number of workers, S is the aggregate years of school completed, Q is quality of schooling, and K is a composite nonlabor input. An individual's education $e_i = Qi$ is simply the product of quantity and quality.

A worker's wage (w_i) (the marginal product), expressed in *logs* is[18]

$$\begin{aligned}(3)\quad log(W_i) &= log(f_1 g_1) + log(1 + Q(g_2/g_1)i) \cong log(f_1 g_1) \\ &\quad + Q(g_2/g_1)i \cong a + b\end{aligned}$$

It consists of the return to the "warm body," $a = f_1 g_1$, and the fractional increment associated with a year of schooling $b = Q(g_2/g_1)$.

Now consider the effect of an increase in quality of schooling on the return to schooling,

$$\frac{dlnb}{dlnQ} = \frac{dlnQ}{dlnQ} + \frac{dln(g_2/g_1)}{dlnQ} = 1 - \frac{1}{\sigma} \tag{4}$$

where σ is the elasticity of substitution between warm bodies and education in $g(\ )$ the labor composite. Improvement in the quality of schooling has two effects on returns: it increases the amount of education associated with given attendance and this effect is to increase b; but, it increases the total stock of education and with diminishing marginal rates of substitution lowers its price. Only if warm bodies and education are good substitutes ($\sigma > 1$) would an increase in quality increase the return to schooling.

In this simple model, there is no life cycle component in the return to schooling. This does not imply, however, that in a single cross section the return to schooling is independent of age (vintage). Suppose that there is steady secular advance in the quality of schooling.[19] In a cross section $b_t = Q(t)g_2/$

[17]Thus, in Table 3, we should compare the schooling coefficient for a 1960 experience group to the experience group in 1970 that contains 10 more years experience.

[18]
$$W_i = \frac{\partial y}{\partial n_i} = f_1(g_1 + Qig_2) = f_1 g_1(1 + Qig_2/g_1)$$

where

$$f_1 = \frac{\partial y}{\partial g}, g_1 = \frac{\partial g}{\partial N}, \text{and } g_2 = \frac{\partial g}{\partial QS}$$

The approximation involves assuming that $Q(g_2/g_1)$ is small, i.e., $ln(1 + bi) \cong bi$.

[19]Let $Q(t)$ represent schooling quality for a cohort entering the work force at time t, with $Q' > 0$ capturing the vintage effect.

TABLE 4—RECENT RETURNS TO EDUCATION: MEAN INCOME RATIOS YEAR-ROUND FULL-TIME WORKERS

	1967		1969		1971		1973	
Age	A	B	A	B	A	B	A	B
25–34	1.34	1.32	1.22	1.39	1.33	1.29	1.30	1.23
35–44	1.38	1.50	1.38	1.54	1.32	1.50	1.21	1.48
45–54	1.31	1.50	1.32	1.65	1.30	1.64	1.37	1.56
55–64	1.26	1.49	1.32	1.66	1.36	1.48	1.27	1.61
65+	1.26	1.44	1.29	1.53	1.28	1.46	1.24	1.40

Source: Current Population Surveys, various issues.
A Income ratio: Males who have completed high school to males who have completed elementary school.
B. Income ratio: Males who have completed college to males who have completed high school.

g_1, but g_2/g_1 depends only on aggregate stocks and is independent of vintage. Therefore $b_2' = Q'(t)g_2/g_1 > 0$ and returns to schooling are higher for more recent cohorts (i.e., younger workers).

For comparing profiles between two cross sections, relative to the earlier profile, $Q(t)$ would have shifted upward as the vintage of the later cross section is increased and g_2/g_1 would have fallen, both because Q has increased and because average schooling levels will themselves have increased. Between cross sections, the return profile can shift upward only if $\sigma > 1$, sufficiently to compensate for the reduction in g_2/g_1 generated by rising average schooling levels. Within cohorts, the life cycle profile should exhibit declining returns as more recent entrants arrive with higher quality and higher average levels of schooling. If this is a reasonable description, then the fact that the returns-experience profiles did not decline for elementary and secondary schooling between 1960 and 1970, and actually increased for college is tentative evidence that schooling's quality increased and that $\sigma > 1$, sufficiently to thwart any downward tendencies associated with rising average schooling levels. Further, the fact that the returns-experience profiles are more steeply inclined for grade schooling for blacks is consistent with a presumption of more rapid secular increases in quality of schooling for blacks. Also, the higher returns to college for the youngest blacks may signal quality improvement, as black students have penetrated traditionally white colleges. The problem is that, aside from the apparent increase in returns to college between 1960 and 1970, the estimated profiles are also consistent with life cycle phenomena. And, the two hypotheses cannot be disentangled without imposing more structure on the model.

The increased return to college attendance that we find between 1960 and 1970 may be surprising. The arrival of the comparatively well-educated postwar baby boom cohorts into the labor market could have signalled an end to the persistently observed high returns to college training. Freeman has in fact argued just this point, except that he presumes that returns to college began to decline after 1969.

Freeman's conclusions are based on comparisons of college/high school graduate earnings ratios between 1969 and 1973, as described in the *Current Population Surveys* (*CPS*). He reports a severe decline in the earnings ratio and argues that since unemployment rates increased between 1969 and 1973, the change occurred despite the business cycle and may therefore be permanent. We have summarized the data for selected years in Table 4.

The most obvious feature of Table 4 is that calendar year effects—the conditions of markets in a particular year—play a larger role in determining skill differentials than most current models presume. Also, if 1969 is taken as a base, as Freeman does, the decline to 1973 looks prodigious. But, if for

example, 1967 is used as the base, the overall decline is much less pronounced. In fact with the exception of the youngest age group, there is little evidence of change between 1967 and 1973.

The fact that cyclical factors may explain declining relative earnings of young workers is added reason for skepticism about the permanency of any decline in the value of higher education. It is generally argued that cyclical downturns offer a relative advantage to more skilled workers. This argument is founded on the presumption of "quasi fixity" or specificity of training on the job being positively correlated with levels of schooling. More skilled workers will be stockpiled or hoarded by firms during periods of reduced labor demand with an eye toward recouping any short-run losses during future expansionary periods. However, if a firm is in the process of hoarding (i.e., underutilizing) its skilled manpower, it surely will not be simultaneously hiring skilled laborers (i.e., recent graduate-entrants). If the theory predicts that skilled workers with job seniority are less vulnerable to cyclic vagaries than others, it must also predict that new entrants to the skilled workforce are more vulnerable than others. The large influx of college graduates that coincided with the recent cyclical downturn seems to have met a predictable fate. Whether their reduced relative wage will persist is uncertain, but the recent experience is a dubious basis either for extreme pessimism or for extrapolation.

B. *Affirmative Action*

Many studies (Link; Weiss-Williamson; Freeman; Joan Haworth, James Gwartney, and Charles Haworth) have assigned a major part of the improvement in relative black income to the effects of government antidiscrimination legislation. Unfortunately, the standard empirical practice is to deduce the impact of the government as a component of the residual – all changes in black-white wage ratios not accounted for by the explanatory variables. It is regrettable that a concept that warrants as much attention as discrimination, is relegated to the "everything else" file in empirical research.

Census data are adequate to identify government's role in changing black-white earnings ratios if one is interested only in the direct effects on those employed by the federal government or in regulated industries because industry of employment is known. Problems arise in identifying effects of government on wages in the private sector. The only method at our disposal was an indirect one – to focus on industries who supply products to governments. For direct employment, we include variables indicating whether the individual is an employee of the federal government and whether he works in an industry that is regulated by the federal government. For those who neither work for the federal government nor work in regulated industries, two additional variables are added. One represents purchases by the federal government as a fraction of value-added originating in the industry. The other is similarly defined for purchases of state and local governments. Executive Orders #11246 and #11375 require that large scale federal contractors comply with the 1964 legislation and establish "affirmative action" plans or risk losing their contracts. We felt that the implied threat of pressures on government contractors for affirmative action gave us our best chance to observe effects of this legislation.

Adjusting for schooling, experience, and location, white 1970 federal employees earn 5 to 8 percent more than others and this discrepancy doubles later in the work career. In 1970, the premium for black federal workers was 10 to 15 percent greater than for whites, but this 1970 wage differential represented a 10 percent drop from that of 1960. In fact, the black-white wage ratio did not change for federal employees between 1960 and 1970. The decline relative to the private sector simply reflects the approximately 10 percent increase achieved in the private sector.[20]

[20]In our constrained estimates the variable for direct employment by the federal government is retained with race and race-year interaction, but year interaction is omitted.

The fraction of all workers employed by the federal government declined slightly between 1960 and 1970. Although blacks are more likely than whites to be federal employees, the proportion of blacks so employed is falling relative to whites and the drop is most pronounced for younger workers. Employees of regulated industries earn 10 to 12 percent more than those in the private sector.[21] Between 1960 and 1970 black employment shares of these industries increased so that regulated industries contributed to rising earnings ratios. The regression coefficients for government's shares of industrial products are very large. They predict for whites that earnings in this form of indirect government employment exceed those of the private sector by one-third to one-half.

We estimated large wage differentials between white employees of federal contractors and those in the private sector. Where whites fare well, blacks appear to do even better. This conforms to our intuition of the effects of affirmative action. The rub is that in these industries implied black-white earnings ratios fall at an average annual rate of 3 to 6 percent per year relative to the private sector (which was rising at about 1 percent per year). The accounting results suggest that none of the government employment variables has an appreciable effect, although the estimated impact of indirect government employment is negative and dominates effects estimated for direct employment (also negative) and for employment in regulated industries (positive).

The Census data indicate that the effects of affirmative action during the 1960's was probably small. Yet, we recognize that these data are far from ideal and we are unable to perform more exacting tests. For example, we did not know whether an individual was employed by a large scale government contractor. And, even if we did, we would want to know much more about the employing firm. How large is it? There should be scale economies since prosecution of a large employer affects more employees. Is it unionized and, if so, what is the union's attitude toward affirmative action? Is it growing, i.e., would increasing the proportion of minority employees require explicit displacement of others? And, most importantly, how dependent is the firm on sales to governments? This final question includes both the government's share of sales and the alternatives available to the firm if the government were not to purchase its product. That is, we expect that defense contractors are much more dependent on governments than, say, are shoe manufacturers *independently of the fraction of a firm's* output of shoes the government happens to buy. The judgment on affirmative action will remain in doubt until these questions are answered. But when our results are matched with those of a number of studies[22] which consider changes in minority employment shares among federal contractors and consistently find only trivial effects, the impression is that the relatively dramatic increases in black earnings achieved during the 1960's is not simply a product of government action. Rather, the evidence is that the most dramatic increases in black relative wages of the 1960's were realized in the most private parts of the private economy.

III. Accounting for Black-White Earnings Differentials

In this section, we present our attempts to account for the black-white wage ratio as it existed in 1970, and for changes in the ratio between 1960 and 1970. Using our *OLS* regression estimates, the *log* of the black-white wage ratio ($\bar{R}$) may be expressed as[23]

$$\ln \bar{R} = (x_1' - x_2')\, b_0 + x_1 \delta_1 \tag{5}$$

The first term on the right-hand side is the main effect of black-white mean character-

[21]Employment in industries regulated by the federal government is included without race, year, or race-year interaction.

[22]See Orley Ashenfelter and James Heckman, George Burman, Morris Goldstein and Robert Smith, and Heckman and Kenneth Wolpin.

[23]For the ratio $\bar{R}$ we have, using equation (1), $\ln \bar{R} = y_1 - y_2 = x_1' B_1 - x_2' B_2 + u_1 - u_2$. The subscripts are: 1 = blacks, 1970; 2 = whites, 1970; 3 = blacks, 1960; 4 = whites, 1960. Since $\delta_1 = B_1 - B_2$ and $b_0 = b_2$, equation (5) follows immediately.

TABLE 5—BLACK-WHITE WEEKLY WAGE RATIOS: OBSERVED RATIOS WITH REGRESSION ACCOUNTING FOR DIFFERENTIALS, 1970

Class and Work Experience	Years of Schooling	Geographic Location	Government Employment[a]	Experience Correction	Total	Residual	*Log* of Observed Weekly Wage Ratio
I: 1-5							
Main Effects	-.175	-.052	-.011	.010	-.228		
Race Interaction	-.491	.077	.013		-.401		
Total	-.666	.025	.002	.010	-.629	.204	-.422
II: 6-10							
Main Effects	-.137	-.039	-.010	.003	-.183		
Race Interaction	-.186	.031	.005		-.150		
Total	-.323	-.008	-.005	.010	-.333	-.107	-.439
III: 11-15							
Main Effects	-.123	-.037	-.011	.001	-.170		
Race Interaction	-.154	-.008	.008		-.154		
Total	-.277	-.045	-.003	.001	-.324	-.157	-.481
IV: 16-20							
Main Effects	-.127	-.038	-.009	.000	-.174		
Race Interaction	-.185	-.027	.019		-.193		
Total	-.312	-.065	.010	.000	-.367	-.123	-.491
V: 21-30							
Main Effects	-.131	-.029	-.008	.000	-.168		
Race Interaction	-.242	-.068	.020		-.290		
Total	-.373	-.097	.012	000	-.458	-.046	-.503
VI: 31-40							
Main Effects	-.139	-.033	-.006	-.001	-.179		
Race Interaction	-.178	-.049	.030		-.197		
Total	-.317	-.082	.024	-.001	-.376	-.137	-.512

[a]Direct; regulated and supply industries

istic differences, weighted by white parameter values, and the second term adjusts for race parameter interaction.[24] The change in the ratio between 1960 and 1970 is written as

$$(6)\quad \Delta ln\,\bar{R} = [(x_1 - x_2)' - (x_3 - x_4)']\,b_0 + (x_1 - x_3)'\,\delta_1 - (x_3 - x_4)'\,\delta_2 - x_3\delta_{12}$$

The main effects of 1960-70 changes in characteristic differences evaluated at 1970 white parameter values is measured by the first term. The second adjusts for race interaction, the third for year interaction, and the fourth for race-year interaction.

In addition to the schooling and government variables described above we have included binary variables indicating residence for the South, North Central, and West regions. Dummy variables are included if the individual resides in a *SMSA* and if the residence is within a central city of an *SMSA*. A variable is also included measuring years in current residence to approximate recency of migration. The remaining class of variables describes a quadratic in years of work experience.

Table 5 summarizes our regression estimates of factors contributing to black-white earnings differentials as of 1970. In all cases, schooling accounts for the largest part of the black-white differential. Except for the first experience class, the schooling effect is approximately equally divided between the main effect of lower grade school coefficients for blacks. For example, in the class with 1-5 years of work experience, the coefficient, -.175, indicates that when

[24]The relative weight given to characteristic differences and coefficient differences is somewhat arbitrary. Characteristic differences could just as easily have been weighted by black coefficients if coefficient differences were weighted by white characteristics. This would have reduced the size of the first term relative to the second.

TABLE 6—AVERAGE ANNUAL PERCENTAGE INCREASE IN BLACK-WHITE WEEKLY WAGE RATIOS, 1960–70: ACCOUNTING ACCORDING TO REGRESSION ESTIMATES BY WORK EXPERIENCE CLASS

Class and Work Experience	Years of Schooling	Geographic Location	Experience Correction	Subtotal	Government Employment[a]	Total	Residual	Observed Increase
I· 1-5								
Main Effects	.91	-.00	.03	.94	.12	1.06		
Interaction:								
Race	-.40	.11		-.29	.03	-.26		
Year	-.22	-.02		-.24		-.24		
Race × Year		1.02		1.02	-.31	.71		
Total	.29	1.11	.03	1.43	-.16	1.27	.96	2.25
II: 6–10								
Main Effects	.76	.10	.01	.87	.05	.92		
Interaction:								
Race	-.19	.12		-.07	-.01	-.08		
Year	-.07	.05		-.02		-.02		
Race × Year		.50		.50	-.21	.29		
Total	.50	.77	.01	1.28	-.17	1.11	.62	1.73
III: 11–15								
Main Effects	.40	.03	.01	.44	-.01	.43		
Interaction:								
Race	-.20	.10		-.10	-.01	-.11		
Year	-.07	.05		-.02		-.02		
Race × Year		.23		.23	-.23	.00		
Total	.13	.41	01	.55	-.25	.30	.75	1.06
IV: 16–20								
Main Effects	.23	.04	-.00	.27	-.00	.27		
Interaction:								
Race	-.26	.14		-.12	.00	-.12		
Year	- 09	-.04		-.13		-.13		
Race × Year				.00	-.33	-.33		
Total	-.12	.14	-.00	.02	-.33	-.31	1.08	0.75
V 21–30								
Main Effects	.46	.06	.00	.52	-.02	.50		
Interaction:								
Race	-.45	.18		-.27	.02	-.25		
Year	-.03	.00		-.03		-.03		
Race × Year				.00	-.34	-.34		
Total	-.02	.24	.00	.22	-.34	-.12	.81	0.70
VI 31–40								
Main Effects	.13	.05	-.01	.17	.01	.18		
Interaction:								
Race	-.34	.15		-.19	.03	-.16		
Year	-.05	-.00		-.05		-.05		
Race × Year				.00	-.30	-.30		
Total	-.26	.20	.01	-.07	-.26	-.33	1.02	0.68

[a]Direct; regulated and supply industries.

weighted by schooling coefficients for whites, the black-white difference in average schooling is large enough to predict black wages (approximately) 17.5 percent below whites. The −.49 is an adjustment for the lower returns blacks gain from schooling.[25]

Even after adjusting for education, experience, and government employment, regional differences in black-white earnings persist. Three characteristics—southern, central city, and metropolitan resident—dominate the geographic accounting in explaining black-white 1970 wage ratios. The southern black wages are the single most important locational source of low black

[25]This rather large racial interaction effect must be considered quite tentative since we find it sensitive to model specification.

relative wages. In the South, white male wages are 8 to 13 percent below those of the Northeast, while black wages in the South range from 15 to 30 percent lower than for blacks in the Northeast.[26] Southern residence reduces the black-white wage ratio from 3 to 13 percent. This differential grows monotonically with experience and mainly reflects different coefficients rather than residence patterns. Declining wage ratios with experience can be attributed either to cohort or life cycle factors and the South may differ from the rest of the country in both. An interpretation that appeals to us is that there are differential vintage effects favoring black southern males for the post-World War II labor market entrants.[27] An alternative explanation is that the presumably more intense discrimination in the South takes the form of restricting blacks from occupations that have rising career wage profiles.

The net effect of all the locational variables is small in the first two experience intervals. In the 11+ experience groups, black wages range from 4 to 9 percent lower because of their locational distribution. The detrimental effect of predominately southern residence is simply much more pronounced for older workers.

As discussed above, the role of government is small. In 1970, the set of government variables should have actually produced black incomes slightly higher than those of whites. The main effect caused by different employment patterns is negative but quite small. The fact that blacks earn more than whites in the federal government and in industries dealing with the government produces the net positive effect.

Table 6 contains our summary accounting for 1960–70 changes in wage ratios. A comparison of Tables 5 and 6 illustrates that one must be careful to distinguish between those factors producing wage differentials at a moment in time and those that cause changes over time. The main effects in Table 6 measure that part of the growth in black-white wage ratios due to contraction or expansion over the decade in black-white differences in characteristics (valued at 1970 white parameter values). The other three terms capture the impact of differential payments for the same characteristics. If blacks earn less than whites for any attribute, the ratio of black to white earnings will decline if this attribute grows secularly. Similarly, the year effect proxies premiums by year. If 1970 was a good year relative to 1960, those individuals with larger amounts of attributes will gain more from the increased "price." Finally, the race-year term is an index of any differential payments blacks received in 1960 relative to 1970 (above the white change between the decade). If such premiums were eliminated over the decade, the black-white wage ratio would fall.

The main effects, those based on changes in characteristic differences, consistently predict rapidly rising wage ratios with schooling playing the leading role. The most rapid increases in schooling occurred in the earlier decades of this century and are therefore more important for those with more work experience, but there are still sizable increases especially for blacks.[28] This large direct effect of schooling is negated

[26]The disparities among the other three regions are less pronounced. In the North Central region for all classes with at least 10 years of experience blacks and whites receive wages 3 to 5 percent higher than the Northeast benchmark. For North Central workers with less than 10 years experience, black wages were higher in 1960 than blacks in the Northeast, but no white wage differentials existed between those two regions. Apparently, black-white earnings ratios increased in both the South and North Central regions relative to other areas. Our estimates suggest that earnings of all persons in the West fell (from 1.4 to 10 percent) between 1960 and 1970 relative to wages in other regions.

[27]Although we rejected year interaction for all experience classes, race-year interaction existed for the three classes with 15 or less years of experience.

[28]The rise in average school completion levels between 1960 and 1970 is as follows:

		Blacks	Whites
Experience class:	1-5	.72	.55
	6–10	1.1	.35
	11–15	1.51	.76
	16–20	1.60	1.02
	21–30	1.87	.98
	31–40	1.52	1.22

mewhat by two factors: the lower coeffi-
nts on black elementary schooling that
duced black relative wages as school com-
etion levels rose both for blacks and
ites, and the 1970 increase in the returns
college where black-white completion
els are high relative to 1960.

Locational effects for those with the least
perience are dominated by race-year
teraction—increased black-white wage
tios rather than a relative movement of
acks toward high wage areas. This is
ainly a result of rising black earnings in
e South and North Central regions be-
een 1960 and 1970.

In sum, our accounting results for sys-
matic determinants of changes in black-
hite ratios are:

1) Geographic location has the largest
nd most favorable effect of factors ex-
mined here. Locational effects are dom-
iated by changed earnings ratios within
gions and migration seems of secondary
nportance.

2) Schooling's role is ambiguous.
lack and white completion levels are con-
erging, but returns to grade schooling are
for blacks than whites. For the first
ee experience classes, with 15 or less
rs of experience, the effect of converging
els is dominant and schooling seems an
portant source of growth in relative black
ome. For those with more than 15 years
experience changed patterns of school
mpletion between the 1960 and 1970
horts result in predictions of falling rela-
e wages for blacks. Black schooling
ns, as measured by increases in number
years, exceed those of whites, but be-
use of differences in returns, the value of
increased schooling of whites (as a pro-
rtion of wages) exceeds the estimated
lue of the increased schooling for blacks.
lative wages are higher for blacks who
e either employed directly by the federal
vernment or are employed in industries
lling a disproportionate part of their
oduct to governments, but the progovern-
ent industrial discrepancy deteriorated
ring the 1960's and we think the catching
p by the private sector is reason for opti-
ism.

IV. Conclusion

We feel the data summarized offer a mildly optimistic picture for the future course of black relative wages. First, the most pronounced gains in earnings ratios are associated with increased schooling levels, and black school completion continues to rise relative to white levels. Secondly, younger cohorts fare better than their older peers. Whatever the cause of the intercohort differentials exhibited in 1970, if the experience of the 1960's is a basis for prediction, wage ratios within cohorts will not decline as time passes, at least, not by very much. Viewed from the historical perspective of relatively constant or deteriorating black-white wages, the experience of the 1960's is encouraging. It is still important to note that although the patterns of gains found between 1960 and 1970 suggest that earnings will rise for blacks relative to whites, the rate of increase is likely to be slow. Among those in the work force both in 1960 and 1970, black relative earnings increased from 0.57 to 0.62. The aggregate ratio in 1970 is roughly 0.64 when the higher earnings of new entrants are taken into account. If the most optimistic view is taken and the relative black wage of new entrants changes by the same amount (11.8) in each new decade as it did between 1960 and 1970 and if the within-cohort growth continues at .04, we will have to wait until the Census of 2000 before parity for new entrants is achieved. Full racial parity would take another 40 years so that few of us will be alive to see it. Since the improvement in the 1960's was exaggerated by the business cycle gains, racial income equality will probably take a good deal longer and could easily be partly nullified for brief periods by a 2 or 3 percentage point increase in the unemployment rate.

REFERENCES

O. Ashenfelter and J. Heckman, "Measuring the Effect of an Antidiscrimination Program," working pap. no. 50, Center Econ. Anal. Human Behavior and Soc. Inst., Nat. Bur. Econ. Res., Stanford 1974.

Gary Becker, *Human Capital*, New York 1964.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.*, July/Aug. 1967, *75*, 352-65.

G. Burman, "The Economics of Discrimination: The Impact of Public Policy," unpublished doctoral dissertation, Univ. Chicago 1973.

R. J. Flanagan, "The Influence of Government Programs on the Relative Economic Gains of Blacks: Marginal Impact vs. Potential Scope," prelim. tech. pap., Office of Assistant Secretary for Policy Evaluation and Research (OASPER), Sept. 1974.

R. B. Freeman, "Changes in the Labor Market for Black Americans, 1948-1972," *Brookings Papers*, Washington 1973, *1*, 67-120.

M. Goldstein and R. Smith, "The Estimated Impact of Antidiscrimination Laws Aimed at Federal Contractors," unpublished tech. pap., OASPER 1975.

G. Hanoch, "Personal Earnings and Investments in Schooling," unpublished doctoral dissertation, Univ. Chicago 1965.

J. G. Haworth, J. Gwartney, and C. Haworth, "Earnings Productivity, and Changes in Employment Discrimination During the 1960's," *Amer. Econ. Rev.*, Mar. 1975, *65*, 158-68.

J. Heckman and K. Wolpin, "Does the Contract Compliance Program Work? An Analysis of Chicago Data," *Ind. Lab. Relat. Rev.*, July 1976, *29*, 544-64.

C. Link, "Black Education, Earnings, and Interregional Migration: A Comment and Some New Evidence," *Amer. Econ. Rev.*, Mar. 1975, *65*, 236-40.

Jacob Mincer, *Schooling, Experience, and Earnings*, New York 1974.

W. Oi, "Labor as a Quasi-Fixed Factor," *J. Polit. Econ.*, Oct. 1962, *70*, 538-55.

S. Rosen, "Short-Run Employment Variation on Class-I Railroads in the Uni States, 1947-1963," *Econometrica*, Ju Oct. 1968, *36*, 511-29.

———, "Towards a Theory of Life Cy Earnings," unpub. paper, 1974.

J. P. Smith and F. Welch, "Black-White Ea ings and Employment, 1960-1970," pa R-1666, Rand Corp., Santa Monica 19

Lester Thurow, *Poverty and Discriminati* Washington 1969.

C. Toro-Vizcarrondo and T. D. Wallace, Test of the Mean Square Error Criteri for Restrictions in Linear Regression *J. Amer. Statist. Assn.*, June 1968, 558-72.

W. Vroman, "Changes in Black Work Relative Earnings: Evidence from t Sixties," unpub. paper, Apr. 1973.

T. D. Wallace, "A Weaker Criteria and Te for Linear Restrictions in Regressio *Econometrica*, July 1972, *40*, 689-98.

L. Weiss and J. G. Williamson, "Black Edu tion, Earnings, and Interregional Mig tion: Some New Evidence," *Amer. Ec Rev.*, June 1972, *62*, 372-83.

——— **and** ———, "Black Educati Earnings, and Interregional Migrati Even Newer Evidence," *Amer. Ec Rev.*, Mar. 1975, *65*, 241-44.

F. Welch, "Labor Market Discriminati An Interpretation of Income Differen in the Rural South," *J. Polit. Econ.*, Ju 1967, *75*, 225-40.

———, (1973a) "Black-White Returns Schooling," *Amer. Econ. Rev.*, Dec. 19 *63*, 893-907.

———, (1973b) "Education and Rac Discrimination," in Orley Ashenfelt and Albert Rees, eds., *Discrimination Labor Markets*, Princeton 1973.

U.S. Bureau of the Census, *Current Populati Surveys*, various issues.

———, "Public Use Samples of Basic Re ords from the 1960 and 1970 Censuses Data Access Descriptions no. 24, Seri CT-8, May 1971.

Tax Credits for Employment Rather Than Investment

By JONATHAN R. KESSELMAN, SAMUEL H. WILLIAMSON, AND ERNST R. BERNDT*

)ver the last generation, the theory and icy of fiscal stabilization have followed ;ely the Keynesian macroeconomics ıch motivated them. From Keynes' origi- formulation down to contemporary tbooks, there has been a mechanical link ween aggregate demand, output, and ployment. Fiscal policy has been viewed ı means of expanding aggregate demand; incentives for investment merely con- trate additional demand in the capital)ds sector. Any substitution between ııtal and labor, while present in neoclas- ıl macro growth models, has typically :n neglected in models of short-run in- ne determination.

›ubstitution towards labor by policy- uced changes in the wage-rental ratio ; virtually been ignored in discussions of :al stabilization. Yet, the implications of . incentives for capital formation have :n analyzed extensively within a neoclas- al framework. This paper will explore the ative employment and investment im- cts of alternative credits in the corporate :ome tax, assuming input supplies to be rfectly elastic. An explicit treatment of : relative unemployment effects is not ›vided here, since it would require the nsideration of supply factors such as)or force participation decisions.[1]

Two job-related policy proposals will be analyzed—an employment tax credit (*ETC*) and a marginal employment tax credit (*METC*). Somewhat analogous to an investment tax credit (*ITC*), the *ETC* would affect the price of labor and might assist in stabilizing the economy. The *METC* departs somewhat from the conventional theory of the firm and therefore will receive close theoretical treatment. The analysis will focus on the factor-substitution effects of the policies and abstract from their expansionary effects on income. Our empirical analysis will contrast *ETC* and *ITC* effects for the aggregate manufacturing sector of the *U.S.* economy between 1962 and 1971. The policy alternatives explored here will be set equal in tax revenue cost to the investment incentives.

I. Background of Employment Tax Credits

A. *Theoretical and Historical Precedents*

Employment tax credits have been advocated previously under the labels "employment subsidies" and "wage subsidies." Most of these proposals have been restricted to the employment of a particular category of worker. With a wage rigidity in the affected region or sector, wage subsidization has been proven superior to capital or output subsidization (see George Borts; Truls Lind and Jan Serck-Hanssen) or to tariff protection (see Jagdish Bhagwati and Vangal Ramaswami). Wage subsidies have been analyzed in a number of contexts: 1) depressed regions of a developed country (see James Buchanan and John Moes; Borts and Jerome Stein); 2) urban sectors of a developing country (see Everett Hagen); 3) income maintenance (see Jonathan Kesselman 1969); and 4) job training of low-wage workers (see Daniel Hamermesh).

Variants of the *ETC* have been directed toward specific sectoral, regional, or demo-

*University of British Columbia; University of wa, University of British Columbia. We gratefully knowledge research assistance by John Lester; help- l comments by the managing editor, a referee, G. ıristopher Archibald, Michael Barth, Daniel Hamer- :sh, A Milton Moore, and David Rose; and financial pport by the UBC Institute of Industrial Relations ıd the University of Wisconsin Institute for Research n Poverty. Opinions expressed are our own.

[1]See Gary Fethke and Samuel Williamson (1976a, b) ›r the analysis of a model with imperfectly elastic upply of labor and unemployment.

graphic groups of workers. In Britain, the Regional Employment Premium and Selective Employment Tax provided labor incentives by location and industry, respectively. Regional *ETC*s including policies which reward firms for raising employment levels have been adopted in Italy, Sweden, Finland, and Germany (see Gösta Rehn). In the United States, *ETC*s with a strong training incentive have been enacted under Job Opportunities in the Business Sector, AFDC Work Incentive Program, and in New York City in the Training Incentive Payments Program. Categorical programs induce firms to substitute eligible types or locations of workers for ineligible ones. Their net effect on economywide employment is thereby muted (see Montek Ahluwalia). Indirect evidence has been found on employment shifting under the AFDC-WIN categorical *ETC* (see Peter Greenston and Duncan MacRae).

In contrast to the categorical approach which allows substitution among categories of labor, an overall employment increase is facilitated by universal coverage. Also, the universal approach avoids contentious political and equity issues that arise with the limitation of eligible groups. A form of universal *ETC* was originally proposed in 1936 by Nicholas Kaldor in a seemingly forgotten paper. Kaldor demonstrated that an *ETC* is the most preferred policy for reducing unemployment under a set of plausible assumptions. A macro-oriented *ETC* concept was further explored by Ragnar Frisch in 1949. In 1977 *ETC*s have become the "new" fiscal tool of the Carter Administration and the U.S. Congress. The Administration proposed an *ETC*, while Congress settled on an *METC*.

B. *Formulation of ETC*

An *ETC* offers the firm a tax credit proportional to some measure of its employment. One type of *ETC* would provide a credit equal to a specified *amount per man-hour* employed.[2] This *ETC* would lower the price of labor to the firm and would als lower the price of unskilled labor relativ to skilled labor. Whether the skilled or th unskilled group gains more of the induce employment depends on the relevant own and cross-price elasticities of demand. Ar alternative type of *ETC* would provide a credit equal to a specified *percentage o wage bill* of the firm.[3] Since this type of *ET* does not change the relative wage rates o the various skill groups, it is more neutra occupationally than the amount per man-hour credit. The percent-of-wage bill approach could be administered with currently reported tax return or social insurance data.[4] The amount-per-man-hour approach would demand the collection of additional information.

One way to achieve greater employment stimulus with the same tax revenues foregone by an *ETC* would be to channel credits to firms for increasing employment Such a formulation will be called a *marginal* employment tax credit (*METC*). Either an amount per man-hour or a percent of wage bill could be applied to the firm's employment increase above a specified base. The firm's base—in man-hours or wage bill could be defined as a percentage of l year's or as a more complex function previous years' magnitudes. The *ME* parallels an *ITC* more closely than does ETC. Both the *METC* and the *ITC* su sidize new purchases of the subsidized put. An asymmetry arises because all un of labor are hired each period, whereas on part of the firm's capital stock is newly pu chased. Thus, investment flow (net of r placement purchases) is analogous to ma

[2]We shall treat the subsidy rate on man-hours or wage bill directly, rather than indirectly through the tax-credit rate. The two are related through the rate of corporate income taxation.

[3]This approach has been called a wage-bill subsi (see Neil Weiner, Robert Lamson, and Henry Peskin *In a third approach called a wage-rate subsidy* (s Kesselman, 1969), the firm's credit is proportional man-hours but inversely related to wage rates. Th method has stronger incentives to employ low sk workers, but, since it requires man-hour records b wage-rate class, is more difficult to administer.

[4]Temporary reduction or elimination of the em ployer contribution to Social Security could provid a convenient way to implement an *ETC*. However to the extent that employers' contributions are shifte onto workers, the efficacy of an *ETC* would b weakened.

inal or additional employment by the irm.[5]

The *METC* rewards the firm when its mployment expands irrespective of the ause for expansion. A similar but less evere procyclical tendency accompanies an ETC. Discretionary applications of an *ETC* or *METC* could reverse the procyclical efect Alternatively, the base or the tax-credit ate could be linked by formula to an aggregate measure of economic utilization, such as the unemployment rate. The *ITC* has had a history of discretionary application and removal, and it is likely that an ETC would be applied in a similar way.

Problems of program definition would accompany an *ETC* or *METC* policy. For example, an *METC* might offer incentives for mergers or fictitious reorganizations of firms. Well-designed rules are needed to avoid such undesirable reactions. A firm acquiring another preexisting operation might be required to include its previous employment (or wage bill) in its own base calculations. In general, though, the measurement of work hours and wage bill is simpler than the measurement of investment especially when the latter needs to be differentiated by equipment and structures For this reason, the problems of program definition may not be as severe as those accompanying investment tax incentives.

II. Theoretical Analysis

The response of a representative firm to ployment tax credits can be analyzed in omparative statics framework. While this y strain the relevance to very short-run bilization policies, it likely captures the ilibrium effects of a sustained policy.[6] is approach is also essential in the analyof investment incentives, where capital ustments are assumed to occur. We assume technology of the firm to be homothetic with positive marginal products and strictly convex isoquants. Because the firm is assumed eventually to encounter diseconomies of scale, its resulting long-run average cost curve is U-shaped. The firm is assumed to be cost minimizing and to face perfectly elastic input supplies.

Analysis of the firm's response to an *ETC* is straightforward. After an exogenous change in effective input prices, the firm chooses a new cost-minimizing mix of inputs for the given output. It follows that average cost net of the credits must be lower in the presence of an *ETC*. Neoclassical analysis of the *ITC* follows a similar procedure.

Unlike the case of the *ETC*, the firm will not always find it advantageous to accept the *METC*.[7] Clearly, a firm will accept the *METC* if its employment without the credit available would have exceeded its current base. The firm may also accept the *METC* if its employment without the credit would have been less than the base. If the base is sufficiently large, the firm will not accept the *METC*. We illustrate the conditions for this to occur with a two-input case; an extension to multiple inputs is straightforward.

In Figure 1, assume that the firm wishes to produce on isoquant V^* at minimum cost. Without loss of generality, costs can be measured in units of capital (K). Assume that C_0D is an isocost curve reflecting the market prices of the inputs. The steeper curve EF reflects the changed input price ratio on *marginal* units of labor (L) receiving the *METC* subsidy. In particular, EF has been constructed tangent to isoquant V^* at S. The intersection of EF and C_0D at point I_0 defines the base Z_0 at which the firm will be indifferent to the *METC*. That is, the kinked schedule C_0I_0F defines an isocost curve for the firm under the *METC*, and it can produce output V^* at a cost of C_0 with or without the credit. With a base of Z_1, less than Z_0, it will minimize its net production costs C_1 by taking the *METC*

[5]See William White for discussion of a marginal inment tax credit

[6]The work of M. Ishaq Nadiri and Sherwin Rosen gests that firms adjust their labor inputs more ckly than their capital inputs. This finding suggests t the *ETC* policies may be preferable to *ITC* icies for short-run stabilization.

[7]The overtime wage subsidy contains analogous issues for a worker's supply behavior under a convex budget constraint (see Kesselman, 1971).

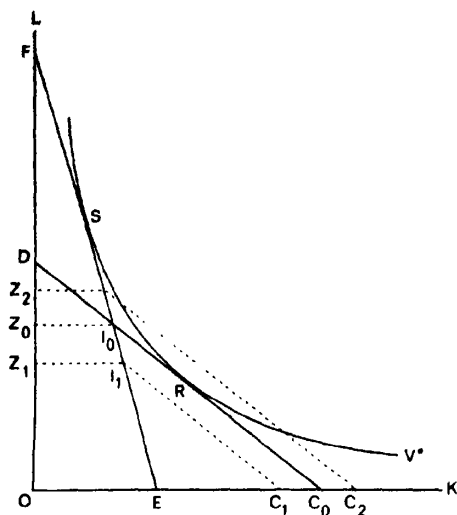



FIGURE 1

and producing with the input mix of point S. In that case, an isocost curve is schedule C_1I_1F. With a base of Z_2, greater than Z_0, the firm will minimize production costs C_0, reject the *METC*, and produce with the input mix at point R.

The first-order conditions for cost minimization are that the firm's choice of inputs will equate relative marginal factor prices with relative marginal physical products. Under an *METC* the marginal factor price for labor has two possible values. A lower price applies to units of labor exceeding the firm's current employment base, which are subsidized, while the higher market price must be paid on labor units below the base. An alternative input criterion would equate relative average factor prices with relative marginal physical products. This, however, would not be consistent with cost minimization in the presence of an *METC*. It is noted that with an *ITC* a potential divergence between average and marginal prices of capital exists. This problem has traditionally been finessed by positing the existence of an active rental market for capital services (see Robert Hall and Dale Jorgensen).

We now compare the firm's cost curves in the absence and presence of the *METC*. All curves assume full adjustment of inp and thus represent long-run cost curv Define the firm's gross cost of inputs market prices and its net cost as gross c minus the *METC* transfers. Define aver cost (AC) and average net cost (ANC) cordingly. Given an output V^* with base Z_0 selected so as to make the firm different to acceptance of the *METC*, t AC must equal ANC at that output. output levels exceeding V^*, ANC must below AC; a proof is straightforward. cause of scale diseconomies at the fi level, the ANC curve is also assumed to U-shaped.[8] The minimum costs on the and ANC curves are denoted P' and P respectively. The associated firm outp levels are V' and V''. These relations seen in Figures 2 and 3.

Before proceeding with the analysis the firm's cost curves, it is helpful to turn the product market. Price competition a the associated adjustment of inputs a entry or exit of firms will drive price to t lower of P' and P''. If P'' is less than the *METC* will be accepted; otherwise will be rejected.[9] With either of these de sions, scale of the individual firm is dete minate.[10] Output price and the market d mand schedule determine the indust aggregate sales. Hence, the number of firn in the industry is also determinate.

The first case to be distinguished appea in Figure 2. Here the AC and ANC curv cross at quantity V^*—which is less than This case reflects directly the choice of th employment base as well as the nature c production technology. Since to the right o V^* the ANC curve must lie below the AC curve, P'' is lower than P'. The firm wi

[8]The depiction of ANC to the left of V^* represen a range of output for which the credit would be r jected. Also note that the ANC schedule does n exist for sufficiently small output levels, for any giv positive employment base.

[9]This rule further requires that the conditi $P' > P''$ must imply the firm's employment exce its base. For a proof that this holds, see Bern Kesselman, and Williamson.

[10]The assumption of constant returns to scale at t firm level would lead to indeterminancy of firm a and other conceptual difficulties in the present co text. See Berndt, Kesselman, and Williamson

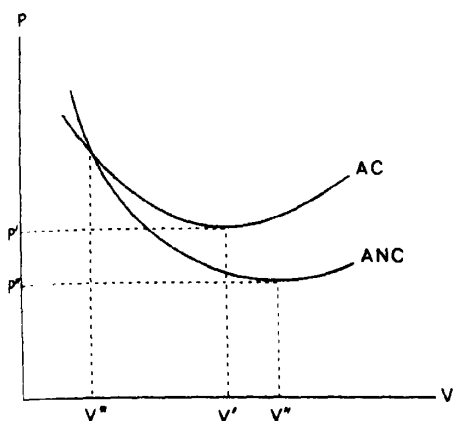



FIGURE 2

-nsequently accept the *METC*. The *C* is merely a special subcase in which */C* falls everywhere below *AC*. It is easily ablished that V'' exceeds V', since the pressing effect of the *METC* on marginal st effects an additional degree of scale economy. Thus, acceptance of the *ETC* will expand firm size. The number firms in the industry will decline unless mand is augmented sufficiently through e fall in output price and other macro-pansionary effects of the policy.

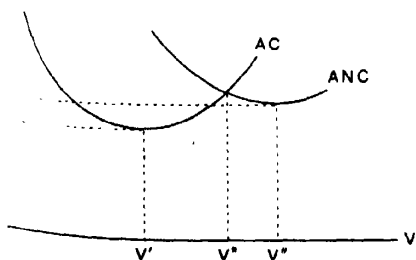
Another set of possibilities arises when larger employment base is selected, so at *AC* and *ANC* cross to the right of V'. this range *AC* is now rising while *ANC* is ill falling. If scale diseconomies are suf-ciently strong, then *ANC* will never fall as w as P'. In this case, P'' exceeds P' and the *METC* is rejected, as illustrated in Figure 3.[11] It is, of course, possible that scale diseconomies are not this strong, so that the *METC* is accepted.

FIGURE 3

A final case is the outcome of P' and P'' equal, which is assigned to rejection of the credit. This captures the effect of a higher employment on future years' base levels of the firm. Other than this borderline case, any dynamic maximizing behavior of the firm is neglected here. When the firm accepts the *METC* in any given year, it does not reckon the impact on future years' credits via its employment base. Such behavior is rational in the presence of uncertainty about and discounting of future events and generous carry-over provisions for unutilized credits.

All input supply functions have been assumed to be perfectly elastic. Further, external economies and diseconomies are assumed not to exist for firms within the industry. The result is a constant-cost industry with horizontal supply curve at price P' or P'', depending upon acceptance of the *METC*. Despite the scale diseconomies which arise at the firm level, it is not inconsistent to characterize technology of the aggregate industry as constant returns to scale.

III. Estimation and Simulation

Any practical assessment of alternative tax policies for employment and investment depends crucially upon the relative magnitudes of their effects. A framework is now developed for simulating the economic effects of *ETC* and *METC* policies. Investment tax incentives are treated in the same framework. For the industry we choose the manufacturing sector of the *U.S.* economy, which accounts for roughly one-quarter of national output. The period 1962 to 1971 is chosen because investment tax credits

[11]We abstract from the adjustment costs of rapid expansion and the effects on the firm's employment base in future years—either of which could induce rejection of the *METC*. Note that if the *METC* base is a constant, the firm can approach its input choice as a static optimization problem. The *ETC* is a special case of this, with base equal to zero.

were introduced in 1962 and suitable data are available.[12]

It would be desirable for the empirical model to incorporate lags in firms' adjustments to desired input levels. Such a procedure involves problems of consistent specification: for example, with given output, restrictions must be imposed on the parameters of the adjustment process. Another problem involves data construction. Conventional capital rental service price formulae are based on the equality in equilibrium between the price of a durable good and the discounted value of its services. In addition, the shadow cost of capital will likely depend on the path of disequilibrium. For these reasons we assume instantaneous adjustment of input demands to their desired levels. Our simulation results must therefore be interpreted as equilibrium values.

Technology in the *U.S.* manufacturing sector is assumed to correspond to a twice-differentiable production function with constant returns to scale. The three inputs entering the production process are production workers (B for blue collar), nonproduction workers (W for white collar), and physical capital (K). Corresponding to such a production function there exists a cost function.

We choose to estimate a cost function rather than a production function for two reasons. With the cost function input prices are right-hand variables, while with the production function input quantities are regressors. It is more reasonable to assume that prices rather than quantities are exogenous; with the cost function, conventional nonsimultaneous equation estimation techniques can be used. Furthermore, the set of derived demand equations used in our simulations is more easily solved with the cost function.

The form of the cost function is specifie as translog:[13]

$$(1)\qquad \ln P = \ln \alpha_0 + \sum_i \alpha_i \ln P_i + \frac{1}{2}\sum_i \sum_j \gamma_{ij} \ln P_i \ln P_j \qquad i,j = B, W, K$$

where P is price of output, the P_i are input prices, and $\gamma_{ij} = \gamma_{ji}$. Assuming zero profits, linear homogeneity in prices, and that input prices are given, we obtain the derived demand equations

$$(2)\qquad \frac{\partial \ln P}{\partial \ln P_i} = \frac{P_i X_i}{PV} = \alpha_i + \sum_j \gamma_{ij} \ln P_j \qquad i,j = B, W, K$$

where X_i is quantity of the ith input, V is output quantity,

$$\sum_i \alpha_i = 1 \text{ and } \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

The output price frontier equation and the derived demand equations (2) estimated by maximum likelihood.[14] Ba on these parameter estimates, we comp Allen partial elasticities of substitut (σ_{ij}) and own-price elasticities (ϵ_{ii}).[15] elasticity estimates are quite stable over sample period. For 1971, the elasticity e mates are .485(σ_{BW}), 1.277(σ_{BK}), −.477(σ_{W} −.338(ϵ_{BB}), −.192(ϵ_{WW}), and −.533(ϵ_{KK}).

The elasticity estimates are of interest their own right. First, we find that capi and white collar workers are complemer while blue collar workers and capital substitutes. The two types of laborers also mildly substitutable.[16] An implicati

[12]A discussion of data sources appears in Berndt and Christensen. We are grateful to the Office of Productivity Analysis, U.S. Bureau of Labor Statistics, and to Laurits Christensen and Dale Jorgensen for furnishing us with updated data to 1971. The revised and updated data are in Berndt, Kesselman, and Williamson.

[13]For further details of the translog cost funct and estimation procedures, see Berndt and Da Wood.

[14]The parameter estimates (asymptotic t-ratios) as follows: $\ln\alpha_0$, 4.308 (535.94); α_B, 6.28 (36 α_W, .076 (5.04); α_K, .296 (44.09); γ_{BB}, .067 (1 γ_{BW}, −.088 (1.79), γ_{BK}, .021 (2.12), γ_{WW}, .156 (3 γ_{WK}, −.068 (−8.54); γ_{KK}, .046 (14.66). The mo tonicity and curvature conditions are satisfied at e observation.

[15]An extended set of elasticity estimates appears Berndt, Kesselman, and Williamson.

[16]These results differ in magnitude from those ported in Berndt and Christensen but agree in si

TABLE 1—SIMULATED EFFECTS OF ELIMINATING *ITC* (percentage changes)

r	ΔP_K	ΔP	ΔB	ΔW	ΔK	$\Delta(B+W)$
2	1.799	.298	.367	−.131	−.985	.231
5	6.978	1.133	1.398	−.482	−3.674	.890
8	8.046	1.269	1.575	−.538	−4.182	1.000
1	5.261	.743	.948	−.349	−2.701	.570

e Δ denotes a percentage change of a variable from its simulated value ($\hat{P}$, $\hat{B}$, $\hat{W}$, and $\hat{K}$) or from its histor- value (P_B, P_W, P_K, and V); P is the index of output price and V is the index of real output. Variables B, W, K are quantities of blue collar labor, white collar labor, and capital inputs, while the respectively subscripted ariables are their net-of-subsidy prices. In these simulations $\Delta P_B = \Delta P_W = 0$.

these estimates is that investment incen- es which lower P_K will induce greater nand for W and reduce demand for B, en the level of output. Thus investment entives favor white collar workers and versely affect blue collar workers.

A check on the reliability of the model is dertaken for the observed period. Based the parameter estimates, we predict the ice of output ($\hat{P}$) by taking the expo- ntial of (1); we predict derived demands forming $\hat{X}_i = (\hat{P}V/P_i)(\hat{\alpha}_i + \hat{\sum}_j\gamma_{ij}lnP_j)$, = B, W, K. The predicted series are then mpared with the historical data. The ean absolute percentage errors are .01 for 83 for $\hat{B}$, .90 for $\hat{W}$, and 1.36 for $\hat{K}$. We nclude that the estimated model simu- ies the historical period with reasonable curacy

IV. Policy Experiments

A. *Investment Tax Credits*

The relative effects of *ITC*, *ETC*, and *IETC* policies can be illustrated through series of simulations. We first determine le effects of removing certain investment ix incentives during 1962 and 1971. A new is computed assuming that there were no ivestment tax credits or increases in depre- iation allowance provisions since 1961. rom here on, we denote these investment icentives by *ITC*. Given the new P_K along ith the historical P_B, P_W, and V, we solve for P, B, W, and K. The estimated tax rev- enue gain from eliminating the *ITC* aver- aged $875 million per year over 1962 to 1971.[17] Accelerated depreciation provisions initiated prior to 1962 are not considered here, although they carry larger costs in tax revenue. In the simulation results reported below, we do not take account of any in- duced multiplier effects on output demand and factor requirements. Replacement of the *ITC* with an equal cost *ETC* or *METC* is unlikely to have multiplier effects unless distributional effects arise.

The results of eliminating the *ITC* for four selected years in the period 1962 to 1971, while holding output constant, appear in Table 1. Total man-hours of labor em- ployed ($B + W$) would have been about 0.7 percent higher over the period. Employ- ment of blue collar workers B would have been about 1.1 percent higher, while em- ployment of white-collar workers W would have fallen about 0.3 percent. Use of capital services K would have decreased about 3 percent, but because of its relatively inelas- tic demand the income to capital would have been more than 2 percent higher. The price of output would have been about 0.8 percent higher. If price elasticity of demand for output (η) is assumed to be unitary, output would have been lower without the investment incentives (−.297, −1.120, −1.253, and −.738 percent in 1962, 1965,

he differences can be attributed to their alternative t of estimating equations (based on a translog pro- uction function) and a different time period 1929 68).

[17]Depreciation allowance provisions were changed in 1964 and 1971, and an *ITC* was introduced in 1962. Our new P_K series follows the procedure employed by Hall and Jorgensen. Further details appear in the ap- pendix of Berndt, Kesselman, and Williamson.

TABLE 2—SIMULATED EFFECTS OF AN EQUAL COST *ETC* OR *METC*
(percentage changes)

Year	ΔP_B	ΔP_W	ΔP	ΔB	ΔW	ΔK	$\Delta(B+W)$
Case 1: $\eta = 0$, amount per man-hour *ETC*							
1962	- 200	-.147	-.152	.048	- 023	-.116	.028
1965	-.815	-.578	-.619	.201	-.097	-.465	.121
1968	-.824	-.572	-.624	.204	-.101	-.467	.121
1971	-.661	-.456	-.504	.155	-.085	-.374	.085
Case 2: $\eta = 0$, percent of wage bill *ETC*							
1962	-.182	-.182	-.152	038	-.013	-.100	.023
1965	-.735	-.735	- 619	155	-.053	-.394	.099
1968	-.737	-.737	-.624	.153	-.052	-.391	.097
1971	-.585	-.585	-.504	.110	-.040	-.305	.066
Case 3: $\eta = 0$, $\mu =$ 5, percent of wage bill *METC*							
1962	-.337	-.377	-.152	.070	-.025	-.185	.044
1965	-1.366	-1 366	- 634	.291	-.097	-.733	186
1968	-1.350	-1.350	-.656	.281	- 096	-.719	178
1971	-1.144	-1.144	-.499	.215	-.079	-.598	.129
Case 4 $\eta = 0$, $\mu = .9$, percent of wage bill *METC*							
1962	-1.064	-1.064	-.155	.224	-.077	-.585	141
1965	-4.339	-4 339	- 631	.954	-.305	-2.337	.615
1968	-4.037	-4.037	-.598	.849	-.298	-2.173	.537
1971	-4.815	-4.815	- 473	.914	-.353	-2.560	.545

Note: In these simulations $\Delta P_K = 0$. For the *METC*, ΔP_B and ΔP_W are percentage changes in the price of n ginal man-hours exceeding the base. Definitions of most symbols appear in the note to Table 1; η is the price e ticity of demand for output; μ is a parameter of the employment base rule in equation (3).

1968, and 1971, respectively). Due to constant returns to scale at the industry level, the effects on factor demands can be computed by adding these values to the corresponding entries in the last four columns of Table 1.

B. *Employment Tax Credits*

We now examine the effects of simultaneously dropping the *ITC* and instituting an *ETC*. The next two simulations set the cost of an *ETC* equal to the estimated revenue cost of the *ITC* for each year. The top two panels in Table 2 report simulations based on two types of *ETC*—an amount per man-hour and a percentage of wage bill. Table 2 shows the effects of enacting each *ETC* alone; combined effects of removing the *ITC* and replacing it with an equal cost *ETC* are obtained by adding the columns in Tables 1 and 2.

The employment gains from instituting an *ETC* alone are very small if output is unchanged. As expected, the amount per man-hour form is more favorable to *B* as against *W* than is the percent of wage form, but the difference is quite sm When output is responsive to price, t ployment expansion is substantially larg the effect of the *ETC* alone for $\eta = -1$ (be approximated by adding the output fects (.151, .611, .616, and .501 percent 1962, 1965, 1968, and 1971) to the fac demand entries in Table 2.

The preceding policy experiments h restricted the *ETC* to equal the tax rever cost of the *ITC* it replaces in each year. T revenue costs were $201 million in 19 $984 million in 1965, $1.265 billion in 19 and $1.115 billion in 1971. Depending up the year and the labor input, the tax-cre rate amounts to between 1/2 and 3-1 cents per man-hour.[18] The effects of m costly *ETC* programs can be approximat by scaling up the effects in Table 2 in p portion to the increased program cost.

[18]In comparison, employer contributions to So Security in 1971 were nearly 20 cents per hour the average worker.

. Marginal Employment Tax Credits

ne *METC* policy channels subsidies to irm for additional employment beyond ise magnitude. Therefore, the same nue cost can finance a larger percentage ge in the price of subsidized units of r in an *METC* than in an *ETC*. This ild provide a stronger substitution ird employment of labor, so long as the "accepts" the *METC*. Our simulations ider only the more easily administered ent of wage bill *METC* formulation.

nce the estimates and simulations rest i aggregate data, the industry is treated firm. This approach necessarily ignores role of scale diseconomies at the firm , which entered the earlier analysis of *TC* acceptance. To implement this ct would require cross-sectional obser-ons on firms. The omission is not seri-in the present context, as we explore narily the economic effects of the ac-ed *METC*. It is noted, however, that aggregation is likely to understate the ie of circumstances in which the *METC* cepted.

nplementation of an *METC* requires definition of a base rule, Z_t. Because base rule involves previous periods' be-ior, time subscripts "t" are introduced. consider a simple rule for an *METC* on e bill where the base in period t is pro-tion μ of the "firm's" gross wage bill in preceding period:

$$Z_t = \mu(P_{B,t-1}B_{t-1} + P_{W,t-1}W_{t-1}) \quad 0 < \mu$$

simulations take the period to be a . each year's base depends on the solu-values of inputs for the previous year.

ercent of wage bill *METC*s appear as es 3 and 4 in Table 2. Each policy has n constrained to carry revenue costs equal hose of the *ITC* in the same year.[19] The cent of wage bill *ETC* is a polar case of percent of wage bill *METC* with $\mu = 0$. effect of increasing μ can be seen by comparing Cases 2, 3, and 4 in Table 2. Clearly, the higher the base parameter μ, the larger is the marginal input-price change affordable from given revenues.[20] With output constant, movement from an *ETC* (Case 2) to an *METC* with $\mu = .5$ (Case 3) roughly doubles the impact on demands for all inputs in each year. Increasing μ from .5 to .9 (Case 4) further triples the *METC* effects on employment. In most years, the *METC* with $\mu = .9$ raises total employment by more than 0.5 percent. When output has unitary price elasticity of demand, the gain in employment in moving from an *ETC* to an *METC* with $\mu = .5$ is more modest.[21] However, the total employment impacts in these cases are larger.

Earlier we noted that the firm will reject (accept) the *METC* based on the decision rule $P'' \geq P'(P'' < P')$. This implies that for the simple base rule (3) and given value of μ, there may be a threshold rate of credit λ at which the firm accepts the *METC*. To illustrate this, we again treat aggregate *U.S.* manufacturing as a single firm and thereby ignore complications of constituent firms growing at different rates. We assume that the 1970 recession had triggered an *METC* policy in 1971. For $\eta = 0$, threshold combinations of parameters (μ, λ) include (.99, .06), (.995, .15), (1.00, .25), and (1.01, .40). These results are of interest, for they suggest that an *METC* could have been effectively utilized even in a year when manufacturing output declined by 2.4 percent.[22]

D. *Output Price Effects*

Our simulated policy experiments carry implications for inflation through changes

[19]The effects of a more costly percent of wage bill *ITC* can be approximated by scaling up the Table 2 its proportionately.

[20]The simulated effects of replacing the *ITC* with an equal cost *METC* can be obtained by adding the corresponding entries in Tables 1 and 2.

[21]The effects can be approximated by adding the output effects to the factor demand effects of Case 3 as explained earlier.

[22]It was observed earlier that the base must exceed the non-*METC* employment before the *METC* will be rejected. Accordingly, we calculate that, while non-*METC* employment in 1971 was 0.968 of 1970 employment, a base of at least 0.99 of 1970 employment is needed for rejection of the credit in 1971.

in the average price of putput (ΔP). Since the simulations do not include monetary behavior or a full macro model, the conclusions are limited to effects on producers' costs. On average, the *ETC* and *METC* policies reduce output price by about 0.5 percent—close to the average annual revenue cost of $875 million divided by the average annual gross production costs (*PV*) of $180 billion. For any given year there is notably little difference in ΔP across program types.

E. *Summary and Extensions*

The policy experiments simulated here portray a hypothetical alternative to the historical experience. Let us assume that an *ETC* or *METC* had been adopted rather than investment tax incentives from 1962 to 1971, with the same revenue cost in each year. Total employment in *U.S.* manufacturing would have been nearly 0.5 percent to more than 1 percent higher in many of the years. All of the increase would have been in blue collar employment, with some offsetting decreases in white collar employment. Use of capital services would have been from 1 to 6 percent lower during the period. The price of output might have been as much as 0.5 percent higher.

The differential effects of instituting various *ETC* and *METC* policies are notable in some cases. There is a modest gain in total employment expansion from specifying an *ETC* as an amount per man-hour rather than as a percent of wage bill. The *METC* can yield much larger employment expansion, especially when the base represents a higher proportion of the current year's wage bill. If output demand is price responsive, these employment effects are enlarged. All of the policies depress output price by nearly the ratio of their revenue costs to total production costs. Replacing the *ITC* with an *ETC* or *METC* affects substantially the aggregate level and distribution of employment.

Our analytical framework has utilized a number of assumptions that might usefully be relaxed in future research. In particular, attention might be directed toward the assumptions of perfectly elastic input su plies, instantaneous adjustment of fact demands, homogeneity of firms, and t absence of macro income-multiplier effec Further empirical work could be bas upon more disaggregated manufactur data and other industries. Finally, discr tionary application of employment ince tives and more complex base rules mer examination for stabilization policy.

REFERENCES

M. S. Ahluwalia, "Taxes, Subsidies, and Em ployment," *Quart. J. Econ.*, Aug. 197 *87*, 393–409.

E. R. Berndt and L. R. Christensen, "Testing f the Existence of a Consistent Aggrega Index of Labor Inputs," *Amer. Eco Rev.*, June 1974, *64*, 391–404.

E. R. Berndt, J. R. Kesselman, and S. I Williamson, "Tax Credits for Employmen Rather Than Investment," Inst. Res o Poverty disc. pap. 279–75, Univ Wi consin-Madison 1975.

E. R. Berndt and D. O. Wood, "Technology Prices, and the Derived Demand fo Energy," *Rev. Econ. Statist.*, Aug. 197 *57*, 259–68.

J. N. Bhagwati and V. K. Ramaswami, "Do mestic Distortions, Tariffs, and th Theory of Optimum Subsidy," *J. Pol Econ.*, Feb. 1963, *71*, 44–50.

George H. Borts, "Criteria for the Evalua of Regional Development Programs. Werner Z. Hirsch, ed., *Regional Acco for Policy Decisions*, Baltimore 1966

——— **and Jerome L. Stein,** *Economic Gr in a Free Market*, New York 1964.

J. M. Buchanan and J. E. Moes, "A Reg Countermeasure to National Wage S dardization," *Amer. Econ. Rev.*, 1960, *50*, 434–38.

G. C. Fethke and S. H. Williamson, (19 "Employment and Price Level Effect a Variable Base Wage Credit," w pap. 76-26, Coll. Bus. Admin., U Iowa 1976.

——— **and** ———, (1976b) *Emplo Tax Credits as a Fiscal Policy Tool*, J Econ. Comm., 94th Cong., Washing 1976.

agnar **Frisch,** *Price-Wage-Tax-Subsidy Policies as Instruments in Maintaining Optimal Employment,* UN Economic and Employment Commission 1949.

M. **Greenston and C. D. MacRae,** "Categorical Wage-Bill Subsidies: Theory and Application," Urban Inst. paper 3603-03, Washington 1973.

E. **Hagen,** "An Economic Justification of Protectionism," *Quart. J. Econ.,* Nov. 1958, *72,* 496-514.

E. **Hall and D. W. Jorgensen,** "Tax Policy and Investment Behavior," *Amer. Econ. Rev.,* June 1967, *57,* 394-414.

niel S. **Hamermesh,** *Economic Aspects of Manpower Training Programs,* Lexington 1971.

N. **Kaldor,** "Wage Subsidies as a Remedy for Unemployment," *J. Polit. Econ.,* Dec. 1936, *44,* 721-42.

I. R. **Kesselman,** "Labor-Supply Effects of Income, Income-Work, and Wage Subsidies," *J. Human Res.,* Summer 1969, *4,* 275-92.

———, "Conditional Subsidies in Income Maintenance," *Western Econ. J.,* Mar. 1971, *9,* 1-20.

T. Lind and J. Serck-Hanssen, "Regional Subsidies on Labour and Capital," *Swedish J. Econ.,* Mar. 1972, *74,* 68-83.

M. I. Nadiri and S. Rosen, "Interrelated Factor Demand Functions," *Amer. Econ. Rev.,* Sept. 1969, *59,* 457-71.

G. Rehn, "Recent Trends in Manpower Policy," *Euro. Yearbook,* 1973, *21,* 82-112.

N. S. Weiner, R. D. Lamson, and H. M. Peskin, "Report on the Feasibility of Estimating the Effects of a National Wage Bill Subsidy," Inst. Defense Analysis pap. P-545, Arlington 1969.

W. H. White, "Illusions in Marginal Investment Subsidy," *Nat. Tax J.,* Dec. 1962, *15,* 26-31.

Weak Invisible Hand Theorems on the Sustainability of Multiproduct Natural Monopoly

By William J. Baumol, Elizabeth E. Bailey, and Robert D. Willig*

This paper investigates the conditions under which a "natural monopoly" can find a set of prices and a set of products that are sustainable against competitive entry. By a *natural monopoly* we mean an industry whose cost function over some given set of products is such that no combination of several firms can produce an industry output vector as cheaply as it can be provided by a single supplier. A *sustainable* vector is a *stationary equilibrium* set of product quantities and prices which does not attract rivals into the industry. Even if a vector is not sustainable, a monopoly may still be able to protect itself from entry by changing its prices whenever and however necessary, in response to any entry that threatens at that moment. But, by definition, only a sustainable vector can prevent entry *and yet remain stationary.*

I. Ramsey Optimality and Sustainable Price-Output Vectors

These concepts lead to three surprising results that constitute the core of our paper. One of these results, due partly to Gerald Faulhaber, asserts:

THEOREM 1: *It is not true that if a single firm has all of the cost advantages of natural monopoly, there must exist a sustainable vector, i.e., equilibrium prices and outputs that are invulnerable to entry. However, the converse does hold; that is, if the monopoly offer[s] no real cost advantage, it will have no sustainable price-output vectors available to it*

Before stating our main and still mor[e] surprising theorems, we first present wha[t] we refer to as the Ramsey rule for Paret[o] optimal pricing under a budget constraint.[1] This proposition is pertinent to a firm which, because of scale economies, would suffer losses if it were to set the prices of it[s] products equal to the corresponding marginal costs. The proposition is usually derived under the assumption that the optimal product set is known; however, we seek both prices *and outputs* that are Pareto optimal subject to the constraint that the firm earns profits π at least equal to the maximal economic profit E permitted by barriers to entry. Necessary (first-order) conditions for a welfare optimum under such a profit constraint are[2]

$$(1)\quad p_i^* - MC_i = -\lambda(MR_i - MC_i) \quad \text{for } y_i^* > 0$$

$$p_i^* - MC_i \leq -\lambda(MR_i - MC_i) \quad \text{for } y_i^* = 0$$

$$\text{with}\quad \pi = E,\quad \lambda \geq 0$$

where p_i, y_i, MC_i, and MR_i are the prices, quantities, marginal costs, and marginal revenues of the industry's set of actual and potential products, N. All *Ramsey-optimal price-output vectors satisfy (1)* by the Kuhn-Tucker theorem.

From the nature of conditions (1), which are relevant only for a multiproduct firm,

*Princeton and New York Universities, Bell Laboratories and New York University, and Bell Laboratories, respectively. This paper was written as part of a study of scale economies and public goods properties of information sponsored by the Division of Science Information of the National Science Foundation. Crucial contributions to the paper were made by John Panzar, Thijs ten Raa, and Dietrich Fischer. Valuable suggestions were also provided by Janusz Ordover, and students in the advanced theory workshop at New York University. This paper reflects the views and assumptions of the authors, not necessarily those of the Center for Applied Economics at New York University or of the Bell System.

[1]We adopt this name since the rule was first der[ived] by Frank Ramsey.

[2]For proof and other important contributions [to] the subject see Marcel Boiteux and Peter Diamond [and] James Mirrlees. For a history and some related for[ms] of the first-order conditions, see Baumol and Da[vid] Bradford.

is clear that our analysis *must* deal with he multiproduct case (which is, in any vent, the only case encountered in reality).

We can now state our main results—the wo basic weak invisible hand theorems for atural monopoly:

HEOREM 2: *Under a set of assumptions iven below, which include a cost function exhibiting both a form of economies of scale nd of complementarity in production, Ramsey-optimal price-output vectors, i.e., vectors which satisfy (1), are sufficient to guarantee ustainability.*[3]

In other words, if the monopoly selects a ector which satisfies conditions (1) for areto optimality, then its profits can be at the maximal level permitted by barriers o entry, and furthermore its virtue will be ewarded by protection from the threat of ntry! The Ramsey-optimal prices will deter ntry of competing firms attempting to arket goods identical with those produced y the monopolist, as long as the potential ntrants have available productive techiques which are no better than the monoolist's. These prices will also deter entry of rms proposing to market other goods in , which may be very close substitutes to he monopolist's.

In general, these Ramsey-optimal vectors ill *not* be the only sustainable price-output ombinations for the monopolist. However e have:

HEOREM 3: *If a monopoly considers a rice-output vector that is not Ramsey-optial, it cannot ascertain whether it is sustainble without global information about the deand and cost functions for its products.*

That is, if the monopoly has access only local information (in the vicinity of its urrent output vector), then the only oututs with which it can be assured of protecon from entry are those that satisfy conditions (1) for Ramsey optimality. Otherwise, the firm may have to know the demand price and cost relationships over an entire region that includes both the origin and the axes.

In sum, our theorems say that the power of the invisible hand has been underestimated. While the literature implies that the invisible hand is potent only under perfect competition, we find that it exercises some sway over monopoly as well. The invisible hand may leave a monopolist vulnerable to potential entry if its pricing decisions are not in accord with the Ramsey rule, or if technology does not provide a sufficiently strong cost advantage to the monopoly form of industry organization. Conversely, if the monopolist has the cost structure and other relations we postulate, and if it selects its price and product set in accord with the Ramsey rule, it will be safe from competitive entry.

Proofs of Theorems 1 and 3 are relatively straightforward and require very little in the way of formal assumptions. However, a full and rigorous proof of Theorem 2 turns out to be relatively complex, and so we will first provide a heuristic argument, postponing a full statement of assumptions and a rigorous demonstration to Appendix A.

II. Sustainability and Subadditivity

Our analysis deals with a monopoly firm which offers to supply a group of products in the set of goods N over which it has a natural monopoly, and announces a vector of prices at which it will sell them. It then undertakes, perhaps because of a legal mandate, to provide whatever output is demanded at those prices. Its productive techniques are known and available to other entrepreneurs, who evaluate the profitability of entering the market and producing some or all of the goods in N. The potential entrant must plan to undercut or match the monopolist's price if he intends to sell any of a product. The entrant is not required to meet all demand, but can instead splinter the market in an arbitrary manner.

The issue in sustainability is whether any potential competitor can expect to enter

[3]But see Faulhaber for an example in which Ramsey icing is not sustainable. This example is based on a st function which does not exhibit the characteriss we will specify.

and earn a profit if the monopoly firm maintains its announced prices.[4] If every price-output combination for a potential entrant will force it to incur losses, we say that the announced monopoly price-output vector is *sustainable*.[5]

DEFINITION 1: The announced prices of a monopolist p^m are *sustainable* if the monopoly is financially viable at these prices $\pi^m \geq 0$, and if no potential entrant can find a marketing plan for which the anticipated economic profits $p^e y^e - C(y^e)$ cover the costs of entry $E(y^e)$. Here, we define the marketing plan of a potential entrant to be a subset A of the product set N, and vectors of prices and quantities p^e_A and y^e_A for the goods in A. We exclude from consideration the possibility that $y^e_A = y^m$, i.e., that the entrant will replicate the monopolist's output vector. For the entrant to make any sales, his prices must not be greater than those of the monopolist,[6] $p^e_A \leq p^m_A$. The maximal sales available to the entrant are given by the demand functions evaluated at the lowest prices offered. Thus $y^e_A \leq Q_A(p^e_A, p^m_{N-A})$, where $N - A$ symbolizes the products not produced by the entrant.

Note that the sustainability concept relates to the possibility of *any* long-run monopoly equilibrium with stable prices. Whether it represents the last in a sequence of moves and countermoves, or merely an initial decision, the set of monopoly prices at that stage is sustainable if and only if no entrant can cover his costs at or below those prices. If every set of prices and outputs is unsustainable, there can be no station monopoly equilibrium because then ei entry will be successful and so the mon oly will come to an end, or the monop must change prices constantly and in petuity to stay one step ahead of poten competitors.[7]

A second crucial concept for our anal which we take to be the defining charac istic of natural monopoly, is that there *cost advantage* to single firm productior the goods in the set N. That is, one firm produce all *given* vectors of outputs m cheaply, in terms of real resources, than any combination of several firms. This c dition, which is referred to as *strict s additivity*,[8] is defined more formally as lows:

DEFINITION 2: A cost function C over the set of products N is (globa strictly *subadditive* if for any k output v tors $y^1, \ldots, y^k$ such that $\Sigma y^i = y$, $y^i \neq$ and $k > 1$, we have

$$(2) \qquad \Sigma C(y^i) > C(y) = C(\Sigma y^i)$$

At first glance it may seem that s additivity and sustainability are equival notions. But Faulhaber, in some ingeni counterexamples, has shown that s additivity is not sufficient to guarantee s tainability.[9] However, we will prove t subadditivity *is necessary* for sustainabil if the technology of the industry is such t production by a single firm is at least as pensive as that by two or more firms, tl there will exist no vector of monopoli prices that can prevent entry into the ind try. Thus, we have the first of our th theorems which describe the workings o (weak) invisible hand in monopoly mark

[4]We might expect such a Nash entry process if, for example, it were recognized that the monopoly would be prevented by a regulator from reacting to the new competition by changing its price structure or product set.

[5]This definition of sustainability is also used in John Panzar and Willig (1977). They discuss some of the economic causes of unsustainability, and some of the implications of sustainability and unsustainability for regulatory policy toward market-structure and entry.

[6]Where the entrant offers a product not supplied by the monopolist there is, of course, no monopoly market price. In this case, the effective monopoly price is the lowest price for that item that reduces its demand to zero.

[7]Irwin Sandberg (1975) has shown that where cost function is subadditive, as it must be under the sumptions of this paper, then in any sequence of restricted price moves and countermoves the si firm can always undercut a competitor in each roun

[8]The notion of subadditivity is used to de natural monopoly both in the mathematical literat (see, e.g., Faulhaber and Sandberg), and in the n mathematical literature (see for example, Alfred Ka p. 123 or Richard Posner, p. 548).

[9]See Faulhaber and also Stephen Littlechild.

FIGURE 1

That is, strict subadditivity is necessary but not sufficient to guarantee the existence of a sustainable price-output vector for the monopolist.

PROOF of Theorem 1:

Since an example is enough to prove that subadditivity does not guarantee sustainability, we offer in Figure 1 a simple counterexample in graphic form, though the result is in no way dependent on the simple form of cost function used. We assume that market demand is satisfied at output y^*. The cost function $OKLNM$ with $LN < OK$ is easily seen to be subadditive over the range depicted. It never costs more than $OK + LN$ for a single firm to produce any output $y \leq y^*$. Any two or more firms turning out all or part of the output will each have to incur the initial fixed cost OK, so that their total cost will be $\geq 2(OK) > OK + LN$, the cost of single firm production. Thus the cost function is subadditive. Yet, because average costs are lower at y' than they are at y^* (slope of OL less than that of OM),[10] a customer group whose quantity demanded is y' will have an incentive to break off and provide that output for itself. Output y^* is thus not sustainable, nor is the situation stable, since output y' does not clear the market.

This completes the proof of insufficiency. To prove necessity, let p^m be any vector of prices at which the monopoly can cover its costs, i.e.,

[10]For a further characterization of the requirements of subadditivity, indicating why patterns such as those shown in Figure 1 can occur, see Baumol (1977)

$$(3a) \qquad p^m y^m \geq C(y^m)$$

The failure of strict subadditivity at the output vector y^m requires the existence of a nontrivial set of output vectors y^i such that

$$(3b) \qquad \Sigma y^i = y^m$$

$$\text{and} \qquad \Sigma C(y^i) \leq C(y^m)$$

so that multifirm production is cheaper. Using (3a) and (3b), we have $\Sigma p^m y^i = p^m y^m \geq C(y^m) \geq \Sigma C(y^i)$, or $\Sigma(p^m y^i - C(y^i)) \geq 0$. It follows immediately that $p^m y^k - C(y^k) \geq 0$ for some k. Thus a competing firm can offer y^k at prices no greater than p^m and at least cover its costs. This constitutes an opportunity for profitable entry, and so the monopoly cannot be sustainable.

It is easy to indicate in intuitive terms why a cost advantage is necessary but not sufficient for sustainability, and that reason also illuminates the nature of the issues involved in the two concepts. Failure of subadditivity essentially requires the existence of at least one way of breaking up the monopoly coalition into a set of subcoalitions in such a way that *each and every* subcoalition can, potentially, benefit from the split. Failure of sustainability, on the other hand, only requires the coalition of consumers to be divisible into subcoalitions in such a way that at least one of them benefits from the split.

III. Entry Costs, Profits, and Sustainability

Our analysis assumes that the monopoly will capture the rents available to it by virtue of its prior position in the market. The rents are the annual equivalent of the discounted present value of the entry costs or barriers to entry facing new firms in the market, and arise from such phenomena as litigation and the need to acquire consumer "good will." While entry costs may vary with the entrant's output vector, so that we must write them as $E(y)$, we would expect that for output sets which are sufficiently large, the entry costs reach and remain at

some maximal value.[11] We suppose that this maximal entry cost E would pertain if the entrant were to supply any output vector that the monopoly firm could itself reasonably consider producing and, in particular, that E holds for any of the optimal outputs satisfying the Ramsey conditions (1).

Under this assumption about entry costs, we immediately have:

LEMMA 1: *If all relevant functions are continuous, any price-output combination p^m, y^m that yields monopoly profits greater than the costs of entry cannot be sustainable. That is, sustainability requires*

$$\pi^m \leq E \tag{4}$$

PROOF:

The derivation of this result is trivial. Suppose (4) is violated at prices p^m not all zero. Then, by continuity, there exist prices slightly lower than p^m (for the nonzero elements p_i^m) at which total revenue minus total operating costs still exceed entry costs E. Therefore another firm can enter and, charging these lower prices to undercut the former monopoly, it can still cover its entry costs.

Thus, E is an upper bound on the profit of a sustainable monopoly.

IV. Pertinent Cost Characteristics

We have seen in Section II that subadditivity of costs is simply not a condition strong enough to guarantee sustainability. Yet, it is a necessary condition. Panzar and Willig (1977) have shown further that it is necessary that the technology have nondecreasing returns to scale at the candidate sustainable vector. Moreover, their analysis demonstrates that for sustainability, product specific economies of scale must be outweighed by the cost savings from joint production (complementarity in production).

There are two cost attributes, devised by Baumol (1977), which together assure cost subadditivity, nondecreasing returns to scale, and production complementarities strong enough to guarantee sustainability. The first of these cost attributes is a strong form of economies of scale. Although it is generally impossible to define average cost for a multiproduct firm, we *can* define the behavior of average cost along a ray in output space, i.e., for proportionate changes in outputs.

DEFINITION 3: A cost function has *strictly decreasing ray average cost*[12] if

$$C(\gamma y) < \gamma C(y) \tag{5}$$

for $\gamma > 1$, for any output vector $y \neq 0$;

that is, if a *proportionate* increase in *all of the firm's outputs* produces a less-than-proportionate increase in its total cost.

The second cost attribute we will be using is a strong form of economies of scope.[13]

DEFINITION 4: A cost function is *transray convex* along a hyperplane $\Sigma w_i y_i = k$, all $w_i > 0$, if given two distinct output vectors y^a and y^b on that hyperplane,[14]

$$C\{\lambda y^a + (1 - \lambda)y^b\} \leq \lambda C(y^a) + (1 - \lambda)C(y^b) \quad \text{for } 0 \leq \lambda \leq 1 \tag{6}$$

Condition (6) asserts that the production cost of a weighted average combination of any pair of output vectors y^a and y^b is not greater than the weighted average of the

[11]Joe Bain, p. 171, seems to imply that such variations are not likely to be substantial. But, in any event, it seems quite plausible that for an entrant to have attained anywhere near the scale of output of the (former) monopolist, he must have incurred *all* the costs that impede entry into the industry. That is all we are assuming here.

[12]This concept is implied by increasing returns to scale (see Panzar and Willig) and by strict ray concavity (see Baumol, 1977). It rules out profitable marginal cost pricing except at points of inflexion in the ray average cost curve. See Baumol (1977) and Panzar and Willig (forthcoming) for an examination of the conditions necessary and sufficient for marginal cost pricing to be unprofitable.

[13]This term was coined by Panzar and Willig (1975).

[14]It can be shown that in order to avoid contradiction between the requirements of (5) and (6) the latter condition can only be postulated to hold for some restricted range of values of w.

FIGURE 2

costs of producing each of them in isolation. This condition requires, in essence, that complementarities in production outweigh scale effects, along the relevant cross section.

Our definitions are readily interpretable graphically. Figure 2 illustrates strictly decreasing ray average costs (5), with the slope of the rays OA, OB, etc. decreasing the further out one is along the fixed-proportions ray OY. As Figure 3 indicates, the complementarity concept (Definition 4) means that along some negatively sloping[15]

FIGURE 3

[15]That the cross section is negatively sloping follows from the requirement that $w_i > 0$.

FIGURE 4

cross section of the diagram, the cost curve is (not necessarily strictly) convex, so that the (unit) cost savings from the production of any weighted-average combination of y^a and y^b outweigh the (unit) cost increases resulting from the losses in economies of specialization.

Thus, a cost surface satisfying Definitions 3 and 4 will have the general appearance of $OCABDE$ in Figure 4. Such a cost surface can be viewed as the natural modification of the class of neoclassical cost functions in which nonincreasing returns are replaced by increasing returns.[16]

V. Geometry of Sustainable Prices

We can now proceed to provide conditions sufficient for sustainability. Since they are readily apparent from the geometry of the situation, we begin with a graphical analysis. For simplicity, we restrict our attention to the case of two outputs, each of which is optimally produced in positive quantity. Further, we suppose that the cost function is that specified by Definitions 3

[16]A cost function which fulfills our conditions can be constructed by beginning with a constant returns cost function which exhibits cost complementarities, i.e., the marginal cost of good i does not increase when the production of good j increases ($C_{ij}(y) \leq 0$, $i \neq j$). It can be shown that such a function is trans-ray convex. If we then append to it a fixed set-up cost, the resulting cost function has strictly decreasing ray average costs and is trans-ray convex, as Definitions 3 and 4 require.

FIGURE 5

and 4. In Appendix A we show that the results hold for the entire natural monopoly product set N, for corner as well as interior solutions, and for less strict assumptions on the cost function.

Figure 5 displays the *total* cost surface, $C(y) + E(y)$. Here our inclusion of entry costs can be considered analogous to the inclusion of normal profits in the cost function for a competitive firm. The particular shape shown for this function is that specified by (5) and (6). Figure 5 also shows a portion of the total revenue surface, the shaded dome $yP(y) = \Sigma y_i P^i(y)$, given by the market's inverse-demand function. The heavy closed curve B is the intersection of the total cost surface with the total revenue surface. Because of its shape we refer to it as a (floating) *hyperbagel.*[17] The projection of the curve B onto the floor of the diagram gives the boundary of the set of outputs T for which the profit of the monopolist, $\pi^m(y)$, is greater than or equal to the entry costs $E(y)$. We assume that T is convex. Further, we shall suppose that region T contains every output vector from which the entrant can hope to earn a profit. This assumption is plausible[18] since we may well expect that the revenue an entrant can earn from the production of any output vector will not exceed the revenue the monopoly could have earned from the sale of the same vector.[19]

We now consider the pricing decision facing a monopolist seeking a set of prices sustainable against entry. The monopolist announces a profitable set of fixed prices $h = (h_1, \ldots, h_n)$ for its n outputs, and he offers to sell as much of his products to his customers as they desire to buy at these prices. Since a potential entrant must plan to set prices at or below the monopolist's, if he is to be able to sell any of the corresponding products, the revenues he anticipates, R^e, are represented by points on or below a hyperplane H, satisfying $R = \Sigma h_i y_i$. That is,

$$R^e(y^e) \leq \Sigma h_i y_i^e \tag{7}$$

While R^e has the form of a revenue function, it does not represent market revenue because the prices h_i are held fixed, while quantities vary freely without regard to demand conditions. Thus, we shall refer to H as a *pseudorevenue hyperplane.*

The position of the hyperplane H with respect to the cost function $C(y) + E(y)$ is critical for sustainability. First, suppose as in Figure 6, that the pseudorevenue hyperplane H_1 cuts through the total cost surface above T. Then, for at least some demand-price relationships, it seems clear that prices h may not be sustainable because a

[17] It may not be apparent at first glance that the somewhat twisted B locus really does have the shape of what the managing editor of this journal has described, felicitously, as "that delectable dessicated doughnut." But, viewed from above, a (slightly deformed) bagel it surely is. Alas, as any *maven* will confirm, bagels are not what they used to be.

[18] The assumption is plausible but it is possible to devise exceptions stemming from demand complementarity between the outputs of the entrant and those of the (former) monopolist. For example, consider an entrant who (profitably) produces electric stoves for which the former monopolist supplies the electricity. If the monopolist were to produce the entrant's output vector composed of stoves but no electricity, he could hardly expect to make any profit, because no one would buy the stoves if no electricity were available. We are able in Appendix A to expand the region of profitable entry to correspond to some greater range of demand complementarity.

[19] That is, if the entrant's revenue is $R^e(y^e) \leq R^m(y^e)$, then if $\pi^m(y^e) = R^m(y^e) - C(y^e) < E(y^e)$, the entrant's net profit from y^e must be negative since $\pi^e(y^e) = R^e(y^e) - C(y^e) - E(y^e)$.

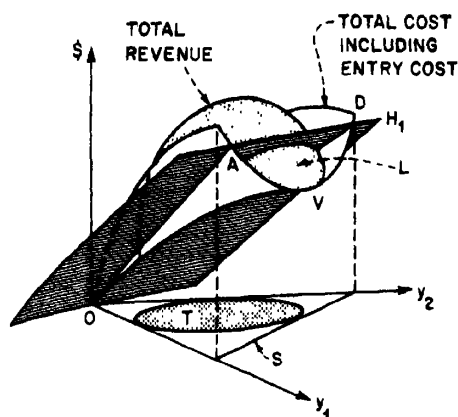



FIGURE 6

FIGURE 7

competitor might enter, offer some output combination in T at prices lower than h (for example, point L), and earn more than the cost of production and entry.

To avoid this problem the monopolist can lower all the prices h_i proportionately so that the hyperplane H swings downward toward the floor of the diagram, until it reaches the lowest hyperplane, H_2 in Figure 7, that still has (at least) one point in common with the hyperbagel B, which we can take to be a point of tangency between H and B.[20] Let the quantities associated with this point be y^m, and the prices defining H_2 be h^m. If these prices happen to satisfy the inverse demand function $h_i^m = P^i(y^m)$, then H_2 is the pseudorevenue hyperplane associated with y^m and pseudorevenue and market revenue coincide at y^m. By construction, H_2 lies below the total cost surface (when it has the shape pictured) over T (except at y^m), and so the sustainability of h^m is assured.

This completes our informal discussion of the geometry of sustainability. We now provide a heuristic proof that an interior *Ramsey-optimum* corresponds to the sustainable tangency point.

VI. A Heuristic Proof of Theorem 2

We first prove that *tangency between H and B is equivalent to a decision by the monopolist to produce at an interior Ramsey-optimal* output vector. For diagrammatic simplicity, we continue to restrict our attention to the two-goods case. Tangency between H and B requires $dH = dC + dE$ at y^* locally along the hypercurve $\pi(y) = E(y)$. Since in the neighborhood of an optimal output vector y^* costs of entry are equal to the constant E, we have $dE = 0$. The requirement $dH = dC$ then yields

$$(8) \quad dH = \Sigma P^i(y^*)dy_i = \Sigma MC_i(y^*)dy_i = dC \qquad i = 1,2$$

while the requirement that dy satisfy $\pi(y) = E(y)$ gives us

$$(9) \quad d\pi = \Sigma\{MR_i(y^*) - MC_i(y^*)\}dy_i = dE = 0 \qquad i = 1,2$$

Solving (8) and (9) for the two products $i = 1,2$, we have

$$(10) \quad \frac{MR_1(y^*) - MC_1(y^*)}{MR_2(y^*) - MC_2(y^*)} = -\frac{dy_2}{dy_1} = \frac{P^1(y^*) - MC_1(y^*)}{P^2(y^*) - MC_2(y^*)}$$

which is clearly the Ramsey-optimality condition (1) for any two products whose outputs are positive.[21]

[20] If the functions are smooth and if $y^m > 0$, then H and B *must* be tangent.

[21] If there are locally increasing returns to scale, the Ramsey rule ((1) or (10)) implies that marginal profit yields are negative, that is, the firm will produce more than it would if it were a profit-maximizing monopoly immune from entry. More than that, (10) may even involve negative marginal *revenues* at the solution point. Thus, Ramsey theory can account for those observed cases in which prices appear to leave the firm in the inelastic portions of its demand curves.

Conversely, suppose $\pi(y^*) = E$, at y^*, $\partial E/\partial y_i = 0$, and that Ramsey condition (10) holds. $\pi = E$ implies that H and $C(y) + E$ coincide at y^*, and (10) implies that (8) holds for dy satisfying (9). Thus, the Ramsey conditions (10) imply that a meeting point of H and B must, under our cost assumptions, be a point of tangency between them.

As a final part of our proof of Theorem 2—that the Ramsey solution is sustainable—we must show that, with costs having the properties we have assumed, the pseudorevenue hyperplane is not only a local support for the cost surface at the Ramsey point y^*, as was just demonstrated, but, as was suggested in the previous section, that it supports costs throughout the potentially profitable region T so that the entrant must lose money anywhere in that region. We can see immediately that this follows from the assumed properties of the cost function, for at y^* the total cost surface "bends away" from H in every direction over T (Figure 7). Thus, at every point in T an entrant's total revenue (which, we know from (7), cannot be above the height of H at that point) must be insufficient to cover his operating and entry costs. Hence, the Ramsey prices will be sustainable, as Theorem 2 asserts.

To show explicitly the role of our cost assumptions, we now state the geometric argument a bit more carefully. Consider any point D in S^- containing T (Figure 8). With the pseudorevenue hyperplane satisfying the Ramsey tangency condition, we want to show that any output D in T must lose money for an entrant because L lies above M, where L and M are, respectively, the corresponding points on the cost surface and pseudorevenue hyperplane H. Draw the cross section $ONKL$ through point L above ray ODN. Then we know that if the cost surface and hyperplane H are tangent at V, at N the cost point K must (by trans-ray convexity), lie on or above point A on hyperplane H. By strictly decreasing ray average cost, point L must lie above the line segment connecting the origin and K, and this in turn must lie on or above the line segment OA containing M. Hence, L must

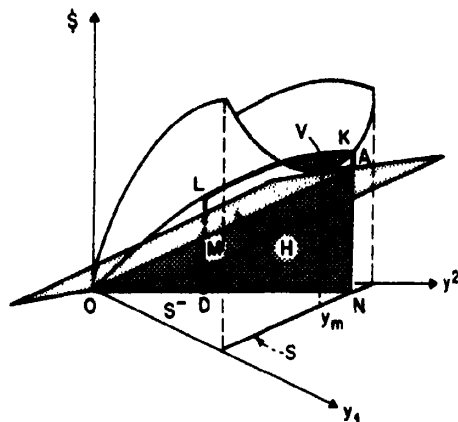



FIGURE 8

lie above M, and so any point D in S^- must be unprofitable to the entrant, which is what Theorem 2 asserts.

VII. Interpretation of Theorem 2

Thus, we have provided a geometric argument showing that Ramsey-optimal pricing offers a reward to the monopolist seeking anticipatory protection from entry. We have, then, a new "invisible hand" notion, applicable outside the well-understood world of perfect competition. Our result is a consequence of the same essential fact that underlies the welfare optimality of perfect competition. With prices held fixed at their market levels, the derivatives of profit with respect to quantities are proportional to the corresponding derivatives of consumer's plus producer's surplus.[22] At the given prices, changes in profit can therefore be used as a local approximation to increments in welfare.

Under perfect competition, producers take prices as parametric because they are so small relative to their markets that they are aware of no influence over prices. They set quantities to maximize profit, given the market prices. Welfare is, fortuitously, also maximized (if functions have the right

[22]Here, as elsewhere in this paper, we restrict attention to one-consumer economies, or to cases in which income is distributed optimally in the eyes of the social decision maker.

shapes) because of the essential fact that parametric-price profit approximates welfare locally.

In our model, the monopolist knows (Lemma 1) that his profits are limited to E. He also knows that potentially viable entry threats are limited to T, and that the profit of an entrant will be less than pseudoprofit (pseudorevenues less costs, including the cost of entry) calculated at the monopoly market prices, held fixed. Hence, if he chooses an output vector with profit equal to the entry cost, which at his fixed prices happens to maximize pseudoprofit over T, then pseudoprofit must be less than entry cost over T. That is, profits of all potential trants must be negative, and the monopst's unique market position is guaranteed stainable.

The monopolist who seeks stationary ices that can protect him from entry, like e perfect competitor, has an incentive to oose outputs which maximize profit *callated at those parametrically fixed market ices*. By doing so, each type of firm inadrtently maximizes net social welfare. ıus, the same invisible hand that guarans welfare-optimal pricing under perfect mpetition, may guide the farsighted mopolist, seeking protection from entry, to e Ramsey welfare optimum.

VIII. Uncertainty of Sustainability without Ramsey Prices

Our final invisible hand theorem (Theon 3) deals with the issue of sustainability non-Ramsey price vectors. It asserts that *any* non-Ramsey prices h (for which H es not support B) the monopolist cannot sure of being able to prevent entry witht knowing the shape of his demand and st functions at outputs that may be far m his current output y^m, and hence far m his range of experience.

ROOF of Theorem 3:

The argument is simple. By the assumed nsupport, there must be some points y^e T at which $\Sigma h_i y_i^e > C(y^e) + E$, as at int L in Figure 6. But, as we have argued (7), the entrant's revenue $R^e(y^e) \leq \Sigma h_i y_i^e$, here the difference $\Sigma h_i y_i^e - R^e(y^e)$ depends upon the magnitude of the inverse demand function at y^e. Hence, at y^e, $R^e(y^e) \lessgtr C(y^e) + E$ and we cannot tell, without knowledge of the demand and cost functions at *every* point y^e in T, whether there will or will not exist points y^e in T at which successful entry is possible.

To discuss Theorem 3 in a more concrete manner, we now present a simple peak load pricing example, and show through it that non-Ramsey price vectors may or may not be sustainable. Let the cost relationship be

$$C(y_1,y_2) = 1 + \max(y_1,y_2) + y_1 + y_2 \quad \text{for } \max(y_1,y_2) > 0$$

which is shaped like a V in cross section.[23] In the cost function, capital cost is $1 + \max(y_1, y_2)$, and each service has a constant unitary marginal operating expense. We suppose that there are no entry costs, and that the inverse demand functions exhibiting independent demands for the two products or services are:

$$P^1(y) = 6 - y_1 \quad \text{and} \quad P^2(y) = 4 - y_2$$

We now pick two candidate non-Ramsey price vectors and show that one is sustainable while the other is not. For our sustainable non-Ramsey solution, we set price just above marginal cost (by the small amount $\delta > 0$) in the off-peak period 2, so that $p_2^m = 1 + \delta$. We then solve the $\pi = 0$ equation for the lowest price in market 1 which just permits the firm to break even.

We first show that these are not Ramsey prices. To see this, note that if we had set $p_2^m = 1$, we would have had $y_2^m = 3$, so that total revenue from that service covers only its variable cost, \$3. Therefore, by continuity, with $p_2^m = 1 + \delta$, the total revenue from market 2 must be $\$3 + \epsilon$ where we can make ϵ as small as we like by selecting δ sufficiently small. This means that if the

[23]This is a version of the cost function usually found in peak-load pricing models (see, e.g., Bailey and Lawrence White). The function is not differentiable everywhere. However, calculations using this function are much simpler to follow than those using a smooth approximation. Note that the function does exhibit transray convexity and ray concavity as our cost assumptions require.

firm is to cover its total cost, market 1 must cover its own variable cost $1 \cdot y_1^m$ plus the capacity cost minus ϵ. This is clearly not possible if its price is set near its marginal cost. Therefore, by (1), these are not Ramsey prices, since $p_1^m > MC_1$, $p_2^m \approx MC_2$. They are, however, sustainable prices. This is so since for any price at or below marginal cost, the entrant will get no contribution to capital cost from his off-peak sales. Hence, he must obtain all capital cost from peak market 1. Since p_1^m is the lowest price that permits the firm to break even in market 1, any $p_1^e \leq p_1^m$ will yield a negative profit if the entrant serves all demand. But because of the fixed component of costs, revenues will be lower than costs for smaller outputs as well. Thus, these non-Ramsey prices of the monopolist are sustainable.

For our non-Ramsey prices that are unsustainable, we set off-peak price so as to equate marginal revenue and marginal cost. Then, $p_2^m = 2.5$, $y_2^m = 1.5$. Solving the $\pi = 0$ equation, peak price and output turn out to be $p_1^m = 4 - \sqrt{21}/2$, $y_1^m = 2 + \sqrt{21}/2$. These are not Ramsey prices because $MR_2 = MC_2$, while $MR_1 < MC_1$. The prices are, however, *not* sustainable, since the entrant can cover his costs at the price output vectors $p_1^e = 1.5$, $y_1^e = 2$, $p_2^e = 2$, $y_2^e = 2$. That is, $\pi^e = (1.5)2 + (2)2 - 1 - 4 - 2 = 0$, where the entrant serves only part of the peak demand. Thus, with the same cost and demand conditions, some candidate non-Ramsey prices are sustainable, while others are not.

IX. Concluding Remarks

We have shown that even under monopoly, the threat of entry can impart some power to the invisible hand. If the monopoly industry has a cost function which exhibits sufficiently strong cost advantages (our strictly decreasing ray average costs and trans-ray convexity), the invisible hand can be shown to encourage a firm to adopt Ramsey prices for its marketed goods, satisfying the relative price conditions (1) for whatever profit level the barriers to entry permit the monopolist to earn. Perhaps still more remarkable (since the issue seems not to have been discussed before in the Ramsey literature), we have shown that this sustainable solution involves production by the monopolist of the Ramsey-optimal *set of goods and services*. For we have been able to establish the surprising result that only this price-output vector can guarantee, independent of any but local knowledge of its demand conditions, to protect the firm against competitive entry.

Our discussion also has one fairly direct implication for public policy. It suggests that the public interest is served by encouraging a monopolist to price in *anticipation* of entry rather than in *response* to it. That is, the monopolist should be encouraged to set prices and outputs that are socially desirable in the first place, knowing that otherwise entry may threaten, rather than changing his prices case by case every time entry seems imminent or actually takes place.

Appendix

A. *A More General Proof of Theorem 2*

While we have provided complete discussions of Theorems 1 and 3, our analysis of Theorem 2—the sustainability of the Ramsey price-output vector—was partially heuristic and unnecessarily restrictive in several respects. The geometric argument assumed that there were only two products, and that positive quantities of *each* were sold by the monopolist. Moreover, the cost function exhibited trans-ray convexity and declining ray average costs, and entry was limited to the potentially profitable region T. We now provide a proof which relaxes each of these assumptions significantly. We begin by providing a complete list of our assumptions and commenting on their new features:

A1 (*Assumption 1*): (a) Potential entrants only consider marketing goods in the natural monopoly product set N. The potential entrants and the monopolist have access to the same production techniques. (b) An entrant incurs entry costs $E(y)$ which attain a maximum value E at "large" output vectors; in particular, $E(y) = E$ for all welfare-optimal (monopoly) output vectors which satisfy the profit constraint $\pi^m(y) \geq \pi^0$, for $0 \leq \pi^0 \leq E$. The profit ceiling E is less than

the unconstrained maximum monopoly profit; i.e., $E < \max\{\pi^m(y) \mid y \geq 0\}$.

A2 (*Assumption 2*): (a) There is a one-to-one relationship between market prices and demand vectors. (b) Market prices depend on the industry output vector, and not on the number of firms that produce it. (c) The cost and inverse demand functions are differentiable over the region $Y = \{y \geq 0 \mid \pi^m(y) \geq E\}$.

A3 (*Assumption 3*): (a) There exists a Ramsey optimum y^* at which (1) is satisfied with $\pi^m(y^*) = E$.[24] (b) y^* lies in $Y^* = \{y \mid \pi^m(y) = E,\ \pi_i^m(y) < 0,\ \forall i\}$, i.e., the outer "northeast" boundary of Y where increases in output decrease profits.

We show in Appendix B that when demands are weak gross substitutes and "normal" in a sense defined there, an optimal Ramsey solution must indeed lie in Y^*. Furthermore, it can be shown that no Ramsey solution outside Y^* is sustainable. Thus, only Ramsey points in Y^* need be considered.

A4 (*Assumption 4*): The net profit function $\pi^m(y) - E(y)$ is strictly quasiconcave over the potentially profitable set, $T = \{y \geq 0 \mid \pi^m(y) - E(y) \geq 0, y \neq 0\}$. More broadly, T is strictly supported at the y^* given by A3 by the tangent hyperplane $S = \{y \geq 0 \mid -\Sigma\pi_i^m(y^*)y_i = -\Sigma\pi_i^m(y^*)y_i^*\}$.

A2–A4 imply that the region which is relevant for our Ramsey optimization is Y, the set where profit equals or exceeds the largest entry cost E. The set T, to which entry threats are limited, does not coincide with Y since, for some y, $E(y) < E$.

A5 (*Assumption 5*): The set T (or, given A3(b) and A4, more broadly, the set $S^- = \{y \geq 0 \mid -\Sigma\pi_i^m(y^*)y_i < -\Sigma\pi_i^m(y^*)y_i^*, y \neq 0\}$) contains every nontrivial output vector from which an entrant can hope to earn a nonnegative net profit.

We show in Appendix B that the set T does contain all possible entrant vectors if demands are weak gross substitutes and "normal." On the other hand, as we have mentioned, limiting entry threats to T is restrictive if there are complementarities in demand. The advantage of using the larger set S^- (the region bounded by the origin and the hyperplane S, whose slopes match those of the outer boundary of T at y^*) is that it permits us to deal with a very broad range of complementarities in demand.[25]

A6 (*Assumption 6*): (a) The total cost function $C(y) + E(y)$ has decreasing ray average costs over $S^- \cup S$. (b) $C(y) + E(y)$ is trans-ray convex on S.

The cost concepts of A6 can be replaced by the somewhat weaker A6′.[26]

A6′ (*Assumption 6′*): The total cost function $C(y) + E(y)$ is strictly supported at y^* by the pseudorevenue hyperplane H above S^-, i.e., $C(y) + E(y) > P(y^*)y$ for $y \in S^-$ and $C(y^*) + E = P(y^*)y^*$.

The virtue of this substitution is that there may be costs which rise sharply with the introduction of a new product. In terms of Figure 5, this means that in practice the cross section of the cost surface above S may first increase and then turn downward as it approaches the axes, thus violating trans-ray convexity. Nevertheless, H may still serve as a support for the cost function above S^- and A6′ permits us to deal with this case.

We now proceed with the proof of Theorem 2. First, we establish the connection be-

[24] Alternately, assume that the set $Y = \{y \mid \pi^m(y) \geq E\}$ is compact and satisfies a Kuhn-Tucker constraint qualification (see Peter Diamond and James Mirrlees, 1971a, b for a discussion of these properties in a similar context). Thus a y^* exists which maximizes welfare over Y, and the Kuhn-Tucker conditions of this program, (1), are satisfied at y^*

[25] An alternative assumption can be shown to yield the result that all y_A^e that are potentially profitable to an entrant must lie in S^-. The assumption asserts that for any potentially profitable y_A^e *either* the former monopolist can find a new vector of quantities which makes both firms viable, or y_A^e is profitable in isolation. This premise amounts to the exclusion of any "destructively parasitic entrant," i.e., one who cannot operate profitably without the presence of the monopolist, but whose presence prevents the survival of the monopoly. The colorful label is inspired by the behavior of such biological villains as the strangler fig or the vine-destroying phylloxera—destructive parasites which simultaneously commit both murder and suicide by strangling the host on which they depend for survival.

[26] See Lester Telser, whose "kind characteristic function" has analogous properties.

tween Ramsey optimality and the points in Y^* at which the pseudorevenue hyperplane supports the total cost surface above S. The argument is a straightforward application of Kuhn-Tucker analysis.

LEMMA 2: *Let $y^* \in Y^*$, and with S defined by A4, let the total cost function, $C(y) + E(y)$, be locally (trans-ray) convex over S at Y^*. Then, y^* satisfies the Ramsey conditions (1) if and only if the pseudorevenue hyperplane defined by $P(Y^*)$ locally supports the total cost surface on S at Y^*; i.e., if and only if*

$$(11) \quad P(y^*)y \leq C(y) + E(y) \quad \text{for} \quad y \in S \cap \mathfrak{N}(y^*)$$

where $\mathfrak{N}(y^*)$ *is some neighborhood of* y^*

and $P(y^*)y^* = C(y^*) + E$

PROOF:

The relations (11) are tautologically equivalent to y^* being a local maximum of the net pseudoprofit function $P(y^*)y - C(y) - E(y)$ over S. Forming the corresponding Lagrangian,

$$L = P(y^*)y - C(y) - E(y) + \mu[\Sigma \pi_i^m(y^*)y_i - \Sigma \pi_i^m(y^*)y_i^*]$$

we obtain the Kuhn-Tucker conditions necessary and sufficient[27] for y^* to be this local maximum of net pseudoprofit over S and to satisfy (11). Since we are in the region over which entry cost $E(y)$ is at a maximal plateau (A1(b)), its derivatives $E_i(y^*)$ must all be zero. Therefore the Kuhn-Tucker conditions are

$$P^i(y^*) - MC_i(y^*) + \mu\pi_i^m(y^*) = 0, \quad \text{for} \quad y_i^* > 0$$

$$P^i(y^*) - MC_i(y^*) + \mu\pi_i^m(y^*) \leq 0, \quad \text{for} \quad y_i^* = 0$$

$$\text{and} \quad y^* \in S$$

Since $\pi_i^m = MR_i - MC_i$ and, by construction, $\pi(y^*) = E$ and $y^* \in S$, these conditions are equivalent to (1).

[27]Note that this is true since total costs are assumed convex locally, so that net pseudoprofit is concave locally at y^*. Further, S is a convex set defined by linear constraints.

We next show that cost assumption A6 implies the support assumption A6′.

LEMMA 3: *Given the entry condition of A1(b) and A3, then A6 implies A6′.*

PROOF:

A1(b) and A3 provide the context needed for Lemma 2 to apply. A6(b) implies that $C(y) + E(y)$ is locally convex over S at y^*. Then, with y^* satisfying (1), Lemma 2 asserts that the pseudorevenue hyperplane supports $C(y) + E(y)$ locally over S at y^* But with the trans-ray convexity of A6(b), a local support is a global support. Hence, $P(y^*)y \leq C(y) + E(y)$, for $y \in S$.

Now, let $y \in S^-$ and define $\hat{y}$ to be its ray extension which lies in S. That is, $\gamma y = \hat{y}$, with $\hat{y} \in S$. Clearly, by the definitions of S and S^-, $\gamma > 1$,

$$P(y^*)y - C(y) - E(y) = \frac{1}{\gamma}[P(y^*)\hat{y} - C(\hat{y}) - E(\hat{y})] + \left[\frac{C(\hat{y}) + E(\hat{y})}{\gamma} - C\left(\frac{\hat{y}}{\gamma}\right) - E\left(\frac{\hat{y}}{\gamma}\right)\right] < 0$$

To see that this expression is indeed negative, note that the first term in square brackets is nonpositive since $\hat{y} \in S$. The second term is negative by declining ray average costs, A6(a) (see (5)). Thus, A6′ is established.

Having shown the relationship between the Ramsey optimality conditions (1) and the condition that the pseudorevenue hyperplane acts as a support for the total cost surface, we finally give our more general proof of Theorem 2.

THEOREM 2: *Given A1–A5, either A6 or A6′ implies that the Ramsey optimum y^* is sustainable.*

PROOF:

Lemma 3 allows us to proceed using A6 Thus $P(y^*) \cdot y - C(y) < E(y)$ for $y \in S^-$. By A5, all nontrivial entry threats lie in S^-. But the anticipated profits of an entrant $p^e y^e - C(y^e)$, are not greater than $p^m y^e$ -

$(y^e) = P(y^*)y^e - C(y^e)$, since $p^e \leq p^m$. ıus $p^e y^e - C(y^e) < E(y^e)$ over S^-, no ıtrant can anticipate covering his entry ısts, and the monopoly is sustainable at $P(y^*)$.

B. *Alternative Set of Assumptions for Weak Invisible-Hand Theorems*

We prove that assumptions A2(a), A3(b), ıd A5 could be dispensed with, if we were ı adopt the premise that the goods in N are ɛak gross substitutes for one another and at demands are "normal," as defined in 8

7 The goods in N are weak gross substitutes. That is, a rise in prices of goods in $-\{i\}$ will never reduce the demand for ɔod i:

$$p^1 \geq p^2, p_i^1 = p_i^2 \rightarrow Q^i(p^1) \geq Q^i(p^2)$$

8(a): (See Sandberg, 1974) Demands for ıe goods ın N are normal, i.e.,

$$p^1 \neq p^2 \rightarrow [Q^i(p^1) - Q^i(p^2)] \cdot (p_i^1 - p_i^2) < 0 \quad \text{for some } i$$

This latter premise asserts that whenever one set of prices is replaced by another, there is at least one good whose demand moves in the opposite direction from its price.[28] If the price of only one product is changed, the response in the quantity demand will be *normal* (its demand curve will have a negative slope). Moreover, no matter what the two sets of prices, the premise precludes demand interdependencies among the various goods so strong as to cause *all* quantities to respond perversely to the price changes.

It is easy to see that with such demands, there is a unique price vector which calls forth each vector of quantities.[29]

LEMMA 4: *If demands for goods are normal (A8(a)), then there is a one-to-one relationship between market prices and demand vectors (A2(a)).*

PROOF:

Suppose the contrary: i.e., $p^1 \neq p^2$ and $Q(p^1) = Q(p^2)$. This immediately contradicts A8(a).

Sandberg has proved that A2(c), A7, and A8(a) together imply that all prices are nonincreasing with quantities. If greater amounts of one or more goods are to be sold, then no prices may rise, and at least one must fall.

LEMMA 5: *(See Sandberg) A2(c), A7, and A8(a) together imply that* $\partial P^i(y)/\partial y_j \leq 0$, $\forall i, j$.

PROOF:

(See Sandberg, 1974) A self-contained proof, provided to us by Thijs ten Raa, is given here. Consider two vectors of outputs, y^1 and y^2. Without loss of generality, we show that when the demand of only one good (good j) increases, then no price increases. Assume

$$(12) \quad y_j^1 < y_j^2 \quad \text{and} \quad y_i^1 = y_i^2, \quad i \neq j$$

Define $K = \{k \mid p_k^1 \geq p_k^2\}$ and $L = \{l \mid p_l^1 < p_l^2\}$. We will show that L is the null set. Define p^3 as follows. For $k \in K$, $p_k^3 = p_k^1$ and for $l \in L, p_l^3 = p_l^2$. Thus, each component of p^3 is the larger of the corresponding components of p^1 and p^2. Since $p_l^3 = p_l^2$ for $l \in L$ and $p_k^3 \geq p_k^2$ for $k \in K$, A7 (weak gross substitutes) implies that

$$(13) \qquad y_l^3 \geq y_l^2, \quad \forall l \in L$$

Suppose L is not empty. Then $p^1 \neq p^3$ and $p^1 \leq p^3$, so A8(a) (normal demands) implies that there is an m with $(p_m^1 - p_m^3)(y_m^1 - y_m^3) < 0$. Since $p_k^1 = p_k^3$ for $k \in K$, m must be in L, with $y_m^1 > y_m^3$. Then, using (13), $y_m^1 > y_m^2$. But this contradicts (12), so it is proven that L is empty, and the result is established. If, moreover, A2(c) is assumed, then $\partial P^i(y)/\partial y_j \leq 0, \forall i, j$.

We need to strengthen this result slightly by assuming that each product's demand curve is downward sloping with no critical points.

[28]It is well known (see Paul Samuelson, for example) that integrable compensated demands have this property for all but price changes that are proportional across the board.

[29]Of course, in the background are fixed outside prices and incomes.

A8(b): At any y, for each i, $\partial P^i(y)/\partial y_i < 0$.

We also need this technical condition:

A9: The Ramsey program of choosing y to maximize welfare over Y admits strict complementary slackness in the Kuhn-Tucker conditions. That is, the necessary conditions (1) can be replaced by the slightly stronger conditions (1′)

$$(1') \quad \begin{aligned} p_i^* - MC_i &= -\lambda(MR_i - MC_i) \text{ for } y_i^* > 0 \\ p_i^* - MC_i &< -\lambda(MR_i - MC_i) \text{ for } y_i^* = 0 \end{aligned}$$

with $\pi = E$ and $\lambda \geq 0$

With these new assumptions, we can establish A3(b).

LEMMA 6: *Under A2(c), A7, A8, and A9, if y^* is the Ramsey optimum given by A3(a), then $\pi_i(y^*) < 0$.*

PROOF:

By definition

$$\pi_i(y^*) = MR_i - MC_i = P^i(y^*) - MC_i(y^*) + \sum_j y_j^* \frac{\partial P^j(y^*)}{\partial y_i}$$

Together with the conditions (1′), this yields

$$(14) \quad \pi_i(y^*) = \frac{1}{1+\lambda} \sum_j y_j^* \frac{\partial P^j(y^*)}{\partial y_i} \quad \text{for } y_i^* > 0$$

$$(15) \quad \pi_i(y^*) < \frac{1}{1+\lambda} \sum y_j^* \frac{\partial P^j(y^*)}{\partial y_i} \quad \text{for } y_i^* = 0$$

Lemma 5 gives us $\partial P^j(y^*)/\partial y_i \leq 0$. Then, (15) immediately yields, recalling that $\lambda \geq 0$, $\pi_i(y^*) < 0$ for $y_i^* = 0$. With $y_k^* > 0$, $y_k^*(\partial P^k(y^*)/\partial y_k) < 0$ by A8(b), and then (14) implies that $\pi_k^*(y^*) < 0$.

It can be seen that this result is surprisingly sensitive to the assumptions. In particular, it is easy to construct well-behaved counterexamples with two complementary goods.[30]

[30]Willig, Baumol, and David Bradford are preparing a paper in which these matters are discussed.

What is most interesting for our present purposes is that assumption A5 can be shown to be unnecessary for demands which are normal and weak gross substitutes. To show this we first prove

LEMMA 7: *A7, A8(a) and continuity of demands imply*

$$(16) \quad \pi^e \equiv p_A^e y_A^e - C(y_A^e) \leq \sum_{i \in A} P^i(y_A^e, 0_{N-A}) y_i^e - C(y_A^e) \equiv \pi^m(y_A^e)$$

where 0_{N-A} symbolizes the vector of zeroes for the products not produced by the entrant. Here (16) asserts, in effect, that π^e, the entrant's profit from the production of y_A^e in the presence of the former monopolist, will never exceed $\pi^m(y_A^e)$, the amount a monopolist could earn by producing the same quantities.

PROOF:

Referring to Definition 1 (of sustainability),

$$\pi^e = p_A^e y_A^e - C(y_A^e), \quad \text{where}$$

$$p_A^e \leq p_A^m \quad \text{and} \quad y_A^e \leq Q_A(p_A^e, p_{N-A}^m)$$

From the definitions,

$$(17) \quad \pi^m(y_A^e) - \pi^e = \sum_{i \in A} y_i^e [P^i(y_A^e, 0_{N-A}) - p_i^e]$$

Note that, by construction, $(y_A^e, 0_{N-A}) \leq Q(p_A^e, p_{N-A}^m)$. Then, under A7, A8(a), and continuity, Sandberg's theorem (the differentiable version is here in Lemma 5) yields

$$P(y_A^e, 0_{N-A}) \geq P[Q(p_A^e, p_{N-A}^m)] = (p_A^e, p_{N-A}^m)$$

In particular, for $i \in A$, $P^i(y_A^e, 0_{N-A}) \geq p_i^e$. Hence, the right-hand side of (17) is nonnegative, and (16) follows.

COROLLARY: *A2(c), A7, and A8(a) imply A5.*

PROOF:

$\pi^e \geq E(y_A^e)$ implies $\pi^m(y_A^e) \geq E(y_A^e)$, by Lemma 7.

Thus, with demands that are normal and for goods that are all weak gross substi-

tutes, an entrant can earn at least as much profit alone in the market as he can in the presence of the former monopolist. However, Lemma 7, like Lemma 6, is very sensitive to its assumptions. For example, in the case of two complementary goods, the single product entrant may well be able to earn far more, when the former monopolist offers the other good, than he can earn in isolation. Fortunately, however, even in such cases, A5 can still hold. Where there are many goods, some substitutes and some complements, (17) will not hold for all entrant output vectors, while A5 may well remain plausible.

Finally, we have

THEOREM 4: *Given A1, A2(b), A2(c), A3(a), A4, A7, A8, and A9, either A6 or A6' implies that the Ramsey optimal prices are sustainable.*

PROOF:

Lemmas 3, 4, 6, and 7 show that these assumptions together imply A1–A6. Thus, Theorem 2 applies.

REFERENCES

E. E. Bailey and L. J. White, "Reversals in Peak and Off-peak Prices," *Bell J. Econ.*, Spring 1974, *5*, 75–92.

Joe S. Bain, *Barriers to New Competition*, Cambridge, Mass. 1965.

W. J. Baumol, "Scale Economies, Average Cost and the Profitability of Marginal Cost Pricing," in Ronald E. Griesson, ed., *Essays in Urban Economics and Public Finance in Honor of William S. Vickrey*, Lexington 1975.

———, "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry," *Amer. Econ. Rev.*, forthcoming 1977.

——— **and D. F. Bradford,** "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

M. Boiteux, "On the Management of Public Monopolies Subject to Budgetary Constraints," *J. Econ. Theory*, Sept. 1971, *3*, 219–40.

P. A. Diamond and J. A. Mirrlees, (1971a) "Optimal Taxation and Public Production I: Production Efficiency," *Amer. Econ. Rev.*, Mar. 1971, *61*, 8–27.

——— **and** ———, (1971b) "Optimal Taxation and Public Production II: Tax Rules," *Amer. Econ. Rev.*, June 1971, *61*, 261–78.

G. R. Faulhaber, "Cross-Subsidization: Pricing in Public Enterprise," *Amer. Econ. Rev.*, Dec. 1975, *65*, 966–77.

Alfred E. Kahn, *The Economics of Regulation*, Vol. II, New York 1971.

S. C. Littlechild, "Common Costs, Fixed Charges, Clubs and Games," *Rev. Econ. Stud.*, Jan. 1975, *42*, 117–24.

J. C. Panzar and R. D. Willig, "Economies of Scale in Multi-Output Production," *Quart. J. Econ.*, forthcoming.

——— **and** ———, "Economies of Scale and Economies of Scope in Multi-Output Production," Bell Lab. working pap., Holmdel 1975.

——— **and** ———, "Free Entry and the Sustainability of Natural Monopoly," *Bell J. Econ.*, Spring 1977, forthcoming.

R. A. Posner, "Natural Monopoly and its Regulation," *Stanford Law Rev.*, Feb. 1969, *21*, 548–643.

F. P. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, *37*, 47–61.

Paul A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

I. W. Sandberg, "On the Mathematical Theory of Interaction in Social Groups," *IEEE Transactions on Systems, Man and Cybernetics*, Sept. 1974, Vol. SMC-4, 432–45.

———, "Two Theorems on a Justification of the Multiservice Regulated Company," *Bell J. Econ.*, Spring 1975, *6*, 346–56.

L. G. Telser, "Necessary and Sufficient Conditions for a Non-empty Core for an Arbitrary Number of Participants with Applications to Location, Capital and Inventory Theory," unpublished manuscript, 1975.

Risk and the Theory of Indexed Bonds

By NISSAN LIVIATAN AND DAVID LEVHARI*

In inflationary times there is a tendency on the part of economists to favor the introduction of bonds linked to the price level of commodities. Usually these proposals relate to *government* issued linked bonds, which are supposed to protect the public against the "inflation tax," or to provide a financial asset in which the public could invest instead of increasing purchases of commodities.[1]

Recently, however, there is a tendency among economists to think that linked bonds could also be introduced in the private capital market for handling inflation risks more efficiently. For example, in a recent article in *Fortune* magazine, Milton Friedman recommends issuance of linked ("indexed") bonds by government and goes on to state that "the arrangements suggested for government borrowing could apply equally to long term borrowing by private enterprises" (p. 174). We may also mention the paper by Marshall Sarnat, who uses a formal portfolio selection model to explain the gain in risk reduction for investors by the introduction of linked bonds.

The foregoing views do not seem to have been adopted by the actual participants in the capital market. In spite of recent marked inflationary trends the important capital markets have not developed a significant free (private enterprise) market for linked bonds. This phenomenon calls for an explanation. Some people may argue that there has not been sufficient time for the capital markets to adapt themselves to the current high rates of inflation, and that an important market for linked bonds will emerge in the future if the current tendencies continue. This, however, is doubtful since even in countries with a long inflationary experience we hardly find a private (as distinct from government) market for linked bonds. We are therefore led to question whether there exist some *fundamental* reasons which may inhibit the linked bonds market to develop to a considerable scale as the value of money becomes increasingly less stable.

In this paper we shall attempt to contribute to the understanding of the foregoing problem. Our approach to the problem is based on the risk aspect. Clearly, when no risk is involved, the anticipated inflation is taken care of by a proper increase in the nominal rate of interest. The linkage of bonds to the general price level is therefore related to *uncertainty* regarding the real value of monetary transactions. It is therefore in terms of risk aversion that we shall try to approach the problem. Our basic variable will accordingly be the *variance* of the value of money and not its expected value.

A basic feature of our approach is to analyze the problem in a context of a market equilibrium. In particular, there is a tendency to consider the advantages of linkage only from the point of view of the lender. However, in a market equilibrium the transaction has also to be satisfactory to the borrower. This constraint will be shown to have far reaching implications.

We shall first try to rationalize the bias that so many economists tend to have in favor of linked bonds. We shall show that this bias probably results from viewing the bond market in isolation from other assets and incomes which are subject to inflation risks. Once this is recognized the condition for dominance of linked over nonlinked bonds no longer holds.

Turning to the more important questions we must admit at the beginning that within the framework of our model we cannot explain the *complete* nonexistence of a market for linked bonds alongside with one for nonlinked bonds. Presumably this explana-

*Professors of economics, Hebrew University, Jerusalem. This research was conducted at the Maurice Falk Institute for Economic Research.

[1]See, e.g., George Leland Bach and Richard A. Musgrave and Richard Goode.

tion requires the introduction of transaction costs and discontinuities of various kinds (such as different interest rates for lenders and borrowers). Since however it is not our purpose to explain the *complete* nonexistence of a market for linked bonds we shall not introduce the above complications. Our purpose is rather to explain, given *some* market for linked bonds, why a *growing* uncertainty concerning the real value of money does not necessarily result in an *increasing* size of the market for linked bonds at the expense of the market of nonlinked bonds. Our fundamental result is that while growing inflationary uncertainty stimulates *demand* for linked bonds it *discourages* at the same time the *supply* of those bonds. (An opposite development occurs with nonlinked bonds.) Thus the key to the paradox seems to lie with the borrowers' issuance of linked bonds.

We shall use throughout a very simple model of an exchange economy and a composite commodity to clarify the ideas. This can serve however as a basis for extensions to more general cases.

I. Dominance of Linked Bonds

Consider an individual consumer with a two-period planning horizon. The individual consumes a single perishable commodity and has a demand for real cash balances. In the first period he has an endowment of y_1 units and consumes c units of the commodity. He also has an endowment of $\overline{M}$ nominal dollars and wishes to hold M dollars, which yields direct utility as a consumer's good. He also spends B_N dollars on nonlinked bonds (B_N is negative for borrowers). Assuming for the moment no linked bonds, his budget constraint for the first period is given by

$$p_1(y_1 - c) + (\overline{M} - M) = B_N \tag{1}$$

where p_1 is the price of the commodity in dollars. Let W denote the dollar value of his wealth in the second period. Then

$$W = M + iB_N + p_2 y_2 \tag{2}$$

where i is *one plus* the nominal interest rate on bonds and p_2 is the price of the commodity in the second period. The real value of W is then

$$\frac{W}{p_2} = w = (m + ib_N)\pi + y_2 \tag{3}$$

where $m = M/p_1$, $b_N = B_N/p_1$, $\pi = p_1/p_2$. Suppose that p_2, and hence π, is the only basic random variable in the system. Then the expected value of w (denoted $\overline{w}$) and its variance (denoted V_w) are given by

$$\begin{aligned} \overline{w} &= (m + ib_N)\overline{\pi} + y_2 \\ V_w &= (m + ib_N)^2 V_\pi \end{aligned} \tag{4}$$

where V_π is the variance of the value of money.

Consider two individuals who engage in a transaction involving nonlinked bonds, i.e., for the lender $b_N > 0$ and for the borrower $b_N < 0$ and of equal absolute amount. If both individuals have no demand for money ($m = 0$) then it can be shown that their nonlinked transaction is dominated by a linked one, assuming both individuals are risk averse and have the same $\overline{\pi}$.

The linked transaction involves lending (borrowing) an amount of B_L dollars and receiving next period $rB_L(p_2/p_1)$ where r is *one plus* the real interest rate and p_2/p_1 represents the linkage factor. The value of w under pure linkage is then

$$w = m\pi + rb_L + y_2 \tag{5}$$

where $b_L = B_L/p_1$. Analogously to (4) we now have

$$\overline{w} = m\overline{\pi} + rb_L + y_2, \qquad V_w = m^2 V_\pi \tag{6}$$

Consider linked and nonlinked transactions as *mutually exclusive* alternatives and let $b = b_N = b_L$. Let us set r to be equal to $i\overline{\pi}$ (the price deflated nominal interest factor). Then, given the nonlinked transaction, the linked one will have the same expected value. However, if $m = 0$ we have $V_w = 0$ under the linked transaction while $V_w = (ib_N)^2 V_\pi > 0$ under the nonlinked one. Applying the mean-variance criterion to w we find that both the lender *and* the borrower will prefer to switch from the nonlinked transaction to the linked one. The foregoing result may also be applied to the nonlinked part of any *mixed* (linked and nonlinked) transaction. Hence there cannot exist a

market for nonlinked bonds, since they are dominated by the linked ones. This may be considered as a possible rationalization of the economists' "bias" towards linked bonds. Thus, linked bonds eliminate unnecessary risk bearing by both lender and borrower.

This result is no longer true when the individuals hold positive amounts of money or expect any nominal incomes in the next period (wages). Denoting the variance of w under nonlinked and linked transaction by V_{w_N} and V_{w_L}, respectively, we find

$$(7) \qquad V_{w_N} - V_{w_L} = ib(2m + ib)V_\pi$$

This is always positive for the lender ($b > 0$) who will therefore prefer the linked transaction (assuming as before $r = i\bar{\pi}$). However, if the *borrower* is still to prefer the linked transaction we need the condition $2m + ib < 0$, or

$$(8) \qquad -b > 2m/i$$

i.e., the absolute value of the loan must be sufficiently in excess of his money holdings and his expected nominal incomes. If however these monetary assets are in excess of his borrowing (by a proper factor) then the borrower will prefer the nonlinked transaction. Thus we have no longer unambiguous dominance of linked bonds.

The economic explanation for this basic difference in the attitudes of lenders and borrowers to linkage is simple. For the lender an increase in the price level reduces both the real value of his monetary holdings and of nonlinked bonds. However, for the *borrower*, while this will still reduce the real value of his monetary holdings, it will at the same time reduce the real burden of his debt, and hence tend to increase his net wealth. Thus for the borrower, nonlinked bonds provide a hedge against inflationary losses of real balances.[2]

It may be noted that the condition $2m + ib > 0$ rules out the possibility of dominance of linked bonds for $r = i\bar{\pi}$ but not necessarily for other $r < i\bar{\pi}$. Indeed, if we set $r < i\bar{\pi}$ then, in some cases we may find a sufficiently low value of r which will induce the borrower to borrow linked and will still leave the lender better off in the linked rather than in the nonlinked alternative. There is, however, no special reason to believe that this will be the case. It is still likely that no r can be found which will make the linked transaction acceptable to both parties involved.

The foregoing analysis may suggest that the more important is the nonlinked sector in the economy, the smaller is the chance of a market for linked bonds to develop to sizeable proportions. It also suggests that if during the inflationary process an increasing proportion of nominal incomes is being "indexed" the chance of a linked bonds market to develop increases. This is based on the observation that in the extreme case where all nominal incomes and money become linked, then the linked bonds must dominate the bond market However, some further analysis shows that using the extreme case of full linkage as an indicator for the effect of *partial* linkage of incomes may not be very useful. (We shall omit the elaboration of this statement.) As a tentative conclusion we suggest that the linked bonds market will gain in relative importance as the proportion of linked incomes becomes sufficiently large.

II. Mixed Portfolio

The dominance of linked bonds when $m = 0$ can also be established directly for a market in which we have both linked and nonlinked bonds. The market for nonlinked bonds becomes active only when we have nonlinked assets (represented by m) in the economy. In fact, the size of the nonlinked assets places a constraint on the equilibrium size of nonlinked borrowing.

When the consumer may hold a mixed portfolio his budget equations become

$$(9) \qquad b_N = y_1 - c + \bar{m} - m - b_L$$
$$w = (m + ib_N)\pi + rb_L + y_2$$

Let us assume the following expected utility function:

$$(10) \qquad U = u(c, m) + f(\bar{w}, V_w)$$

which is based on the mean ($\bar{w}$) – variance (V_w) approach. The first-order partial deriva-

[2] A similar point has been made in C. G. Fane.

tives satisfy u_c, u_m, $f_{\bar{\pi}} > 0$ and $f_{V_w} < 0$, where u_c is the partial derivative of u with respect to c and similarly for the other partials. The first-order conditions with respect to the three independent variables (c, m, b_L) can be rearranged to read as follows (see the Appendix):

$$(11) \quad \frac{u_c}{f_w} = r$$

$$(12) \quad \frac{u_c}{u_m} = \frac{i}{i-1}$$

$$(13) \quad i\bar{\pi} - r = S(m + ib_N);$$

$$S = 2\left(-\frac{f_{V_w}}{f_{\bar{\pi}}}\right)V_\pi i > 0$$

The first two tangency conditions have an obvious interpretation. The last condition is obtained from the following:

(13′)

$$\frac{\partial U}{\partial b_L} = f_{\bar{\pi}}(r - i\bar{\pi}) - 2f_{V_w}(m + ib_N)V_\pi i = 0$$

where c and m are held constant and b_N is considered as a function of b_L by (9). Equation (13) shows that if $m = 0$, the individual will always *lend* nonlinked if there is a premium on nonlinked bonds, i.e., if $i\bar{\pi} > r$. If $m > 0$ then the individual may also *borrow* nonlinked under the foregoing condition but his borrowing can never exceed his monetary holdings. Thus it is the hedging principle which governs nonlinked borrowing.

We can first reaffirm the result obtained earlier concerning the disappearance of the market for nonlinked bonds when $m = 0$ for all consumers. For in this case, assuming the same $\bar{\pi}$ for everybody, we have by (13)

$$\text{sign}(i\bar{\pi} - r) = \text{sign}\, b_N$$

Hence, b_N has the same sign for *everybody*. However, in a market equilibrium lenders must be matched by borrowers. Consequently, a market equilibrium can exist only with $b_N = 0$ for everybody and $i\bar{\pi} - r = 0$.[3]

[3]A somewhat similar result has been obtained in a recent paper by Stanley Fischer. One may use our approach for an intuitive interpretation of his result. Note, however, that if $\bar{\pi}$ differs among individuals we can have a market for nonlinked bonds even if $m = 0$.

If however $m > 0$ then

$$\text{sign}(i\bar{\pi} - r) = \text{sign}(m + ib_N)$$

In this case we cannot have a negative risk premium for nonlinked bonds, $i\bar{\pi} - r < 0$, because then $b_N < 0$ for everybody (i.e., everybody wants to *borrow* nonlinked) and no market equilibrium exists. The same is true for $i\bar{\pi} - r = 0$. The only possibility which is consistent with both positive and negative b_N, and therefore with market equilibrium, is a *positive* risk premium, i.e., $i\bar{\pi} - r > 0$.

It can be seen that the amount of nonlinked borrowing is related to the monetary assets. Since $m + ib_N$ is positive in equilibrium, we have $-b_N < m/i$. Any borrowing in excess of m/i will necessarily involve linked bonds.

The foregoing condition shows that the size of monetary assets and incomes places an upper bound on the amount of nonlinked borrowing. This suggests again that the importance of the nonlinked bond sector is related to the weight of nominal (nonlinked) assets in the economy. We should remind the reader that for the purposes of the present analysis m should include not only cash balances but all nonlinked incomes expected in the next period.

III. Unstable Value of Money and the Volume of Linked Bonds

We have seen that instability in the real value of money does not imply any sort of dominance of linked over nonlinked bonds. Is it however true that as the instability of value of money *increases* the market will exhibit an increasing tendency to shift from nonlinked to linked bonds? To analyze this question, we shall consider the effect of an increase in V_π on the volume and the value of transactions in linked and nonlinked bond markets. We take V_π as our independent variable because it is the unanticipated variation in π which is at the heart of the linking procedure.

Let us deal with a mixed market (linked and nonlinked bonds) and consider the effect of increasing V_π on the volume of transactions in the linked and nonlinked markets. We can approach this problem by

working out the reactions of the individual consumer to this change. To simplify the analysis we shall assume $u_{cm} \geq 0$ and a quadratic f in (10):

$$(14) \quad f(\overline{w}, V_w) = \alpha\overline{w} - \beta(\overline{w}^2 + V_w)$$

with $\alpha, \beta > 0$, and $\alpha - 2\beta\overline{w} > 0$. By differentiating the individual's equilibrium conditions (11)-(13) with respect to V_π, we obtain the following changes in the optimal values of c, m, b_L, and b_N (one of these derivatives can be obtained from the others by (9)):[4]

$$(15) \quad \frac{db_L}{dV_\pi} > 0, \quad \frac{dm}{dV_\pi} < 0, \quad \frac{dc}{dV_\pi} < 0$$

$$(15') \quad \frac{db_N}{dV_\pi} < 0 \quad \text{if} \quad r \geq \overline{\pi}$$

Thus an increase in the variance of the value of money increases the demand for linked bonds at the expense of all other decision variables. All this is based on the assumption that $m + ib_N > 0$ which is required for market equilibrium as we saw earlier. In fact, reversing the sign of $m + ib_N$ would reverse the sign of the derivatives in (15). We also used the condition $i\overline{\pi} > r$ which is again derived from market equilibrium. This latter condition is however not required for the result $db_L/dV_\pi > 0$ or for $db_N/dV_\pi < 0$. As for db_N/dV_π the condition $r \geq \overline{\pi}$ is sufficient but not necessary. This condition will certainly be satisfied in inflation ($\overline{\pi} = E(p_1/p_2) < 1$) if the real rate of interest ($r - 1$) is nonnegative. Although in a model without production the latter condition is not necessarily satisfied, it seems still reasonable to suppose that $\overline{\pi} < r$ holds.

The main fact which we wish to emphasize with respect to $db_L/dV_\pi > 0$ (as well as $db_N/dV_\pi < 0$) is that it holds for both lender and borrower of linked bonds. Thus, while for the lender ($b_L > 0$) there will be a tendency to lend more when V_π increases there will be at the same time a tendency for the borrower ($b_L < 0$) to borrow less. Since the supply of linked bonds shrinks at the same time as demand expands, it is not clear at all what the effect on the volume of transactions will be.

The economic interpretation of this fundamental result is as follows. An increase in V_π increases the riskiness of the net nonlinked financial assets ($m + ib_N$). Since these are positive for both lender and borrower in the linked market, an increase in V_π will induce them both to *reduce* their $m + ib_N$ and channel the freed resources into linked bonds. For the linked lender this implies additional lending (increased demand for linked bonds) while for the linked borrower the reduction in ($m + ib_N$) is used to reduce his linked debt (decreased supply of linked bonds). Indeed, if we look at (13') we see that an increase in V_π increases the marginal expected utility of b_L if $m + ib_N > 0$, leading to increased demand and decreased supply of linked bonds.[5]

In order to appreciate this result, we may ask if there is any reason at all to expect supply of linked bonds to *increase* as V_π increases? An argument which points in this direction is the following one. Suppose that there are many borrowers who borrow both linked and nonlinked. Now as V_π increases, their nonlinked debt becomes more risky in real terms. Consequently, the borrower will tend to reallocate his borrowing so as to increase the proportion of linked at the expense of nonlinked borrowing. This implies that an increase in V_π will *increase* the *supply* of linked bonds, contrary to the conclusion derived from our model.

What is wrong with the foregoing argument? The main point which is wrong is that the individual does not consider the riskiness of nonlinked borrowing as such but rather the riskiness of his *total* nonlinked assets and liabilities which are given by $m + ib_N$. If $m + ib_N$ were negative, the foregoing argument would still apply. However, market equilibrium imposes the requirement that $m + ib_N$ be positive even for $b_N < 0$. The increased V_π calls therefore for a reduction in a positive $m + ib_N$ which, given m, implies *increased* rather than re-

[4]For the derivation, see the Appendix. It may be noted that the result $dc/dV_\pi < 0$ is due to the existence of linked bonds. If only nonlinked bonds were allowed the effect on c would be reversed.

[5]Note that $f_{\overline{w}}$ and f_{V_w} are independent of V_π under assumption (14).

uced nonlinked borrowing. This will reuce the need for linked borrowing, thus ausing a reduction in the supply of linked onds.

Turning to the nonlinked market we find gain that an increase in V_π results in counracting tendencies. As the increase in V_π duces a reduction in $m = ib_N$, which is ositive, individuals will reduce their deand and at the same time increase their upply of nonlinked bonds. Again, in order appreciate the role played by the market quilibrium constraint in the foregoing nalysis, suppose *hypothetically* that for onlinked *borrowers* $m + ib_N < 0$. An increase in V_π will then induce them to reduce e absolute value of $m + ib_N$, which (given) implies a reduction in their nonlinked orrowing, i.e., a reduction in the supply of onlinked bonds. Since for lenders in the onlinked market an increase in V_π reduces emand for nonlinked bonds, we find that oth supply and demand for nonlinked onds decrease leading to contraction of the olume of transactions in this market. However, this state of affairs is ruled out by the arket equilibrium constraint which reuires $m + ib_N$ to be positive for lender and orrower alike. We are thus left inevitably ith the counteracting tendencies of deand and supply in the market for nonnked bonds.

Let us illustrate our analysis of the effect f V_π on the market for linked bonds. Dene a unit of linked bonds as an obligation o pay in the next period the value of one nit of the commodity (here we assume for implicity $p_1 = 1$). The real market price of his is $1/r$.[6] If e_L denotes the number of inked bonds then $b_L = (1/r)e_L$. In Figure 1 we draw the excess demand curve of he consumer for e_L, denoted EE' for $< i\pi$. Defining negative excess demand as "supply" we may shift the AE part of EE' o the positive quadrant and consider AS as the individuals' supply curve. An increase in V_π increase b_L for all $r < i\pi$. Hence, rb_L increases at any r and EE' shifts

[6]This equals the marginal future expected utility of one safe unit of the commodity divided by the marginal utility of current consumption, as can be inferred from (11).

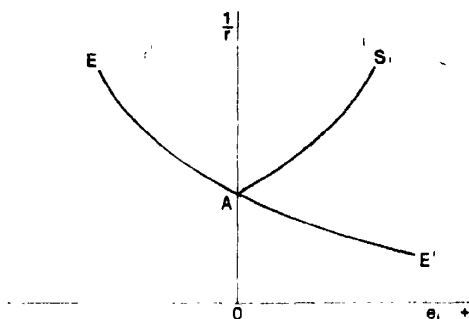



FIGURE 1

to the right. However, a shift of EE' to the right at any r implies a shift of all individuals' demand curves to the right and at the same time a shift of all their supply curves to the left. Thus the *market* demand curve shifts to the right and the supply curve shifts to the left, which may leave the market equilibrium volume unchanged (this refers to $r < i\pi$, the relevant range for market equilibria).

The foregoing analysis indicates there will be a tendency for the price of linked bonds $1/r$ to increase as a result of an increase in V_π. If the shifts in the market demand and supply curves for linked bonds neutralize each other then the volume of transactions will tend to remain constant while its value ($|b_L|$) will increase approximately proportionally to the increase in $1/r$. Alternatively, the value of linked transactions will tend to remain constant in terms of next period's resources (e_L) while it will increase in terms of current resources. On the whole, we can state that the tendency to shift to the linked market as a result of an increase in V_π is limited because of an absence of a definite change in the volume of these transactions.

IV. The Source of Inflation and the Market for Bonds

Throughout our analysis we treated the expected price changes as being exogenous, while from a macro-economic point of view the price variation itself is caused by more fundamental factors. It is therefore necessary to examine the relationship (if any) between the source of inflation and the ten-

dency to link bonds. As will be shown below, there is a basic distinction between the two main types of inflation—a purely monetary inflation and an inflation which is intended to tax the real resources of the private sector.

Let us consider first a purely monetary inflation where the government prints money to make transfer payments to individuals. These transfer payments constitute the basic exogenous random variable of the system. Let us return to the model of Section III and extend it by adding a random nominal transfer payment T_j to the next period's income of individual j. The individual's real wealth in the next period is then

$$(16) \quad w_j = (m_j + ib_{N_j})\pi + t_j\pi + rb_{L_j} + y_{2j}$$

where $t_j = T_j/p_1$.

Now from a macro-economic point of view, the payments T_j in the economy as a whole will of course affect p_2, and hence also π. Let us take this relationship into account by adopting a quantity theory approach. In particular let us assume that the value of *real* balances in the economy in the second period (λ) is a constant[7] (the amount of goods in the second period being constant) independent of the distribution of individual T_j

$$(17) \qquad \frac{\Sigma \overline{M}_j + \Sigma T_j}{p_2} = \lambda$$

where the summation is over all individuals $j = 1, \ldots, n$. We can write (17) alternatively as

$$(18) \qquad \sum_j t_j\pi = \lambda - \pi\sum \overline{m}_j$$

Define the regression coefficient

$$A_j = \frac{Cov(t_j\pi, \pi)}{V_\pi}$$

Then summing over all individuals and using (19) we obtain

$$(19) \quad \sum A_j = \frac{\sum_j Cov(t_j\pi, \pi)}{V_\pi} = \frac{Cov\left(\sum t_j\pi, \pi\right)}{V_\pi} = -\sum \overline{m}_j$$

Thus on the average (over individuals) the real value of transfer payments must vary inversely to the real value of money in order to keep the aggregate value of real balances constant. From (16) we obtain

$$(20) \quad V_{w_j} = (m_j + ib_{N_j})^2 V_\pi + Var(t_j\pi) + 2(m_j + ib_{N_j})Cov(t_j\pi, \pi)$$

Hence

$$(21) \quad \frac{\partial V_{w_j}}{\partial b_{L_j}} = -2iV_\pi(m_j + A_j + ib_{N_j})$$

The modified first-order condition of optimality with respect to b_L can then be written analogously to (13) as

$$(22) \quad i\overline{\pi} - r = S_j(m_j + A_j + ib_{N_j})$$

where $S_j > 0$.

We shall now show that if price expectations are determined according to the quantity theory assumption (17) then in a market equilibrium the risk premium $i\overline{\pi} - r$ is zero. This is a major difference compared with our earlier analysis where the premium was positive.

To prove the foregoing proposition suppose, to the contrary, that $i\overline{\pi} - r > 0$ Then, since $S_j > 0$, we have $m_j + A_j + ib_{N_j} > 0$, for all j. Summing over j we obtain

$$(23) \qquad \Sigma m_j + \Sigma A_j + i\Sigma b_{N_j} > 0$$

But $\Sigma A_j = -\Sigma \overline{m}_j$ by (19) and in a market equilibrium $\Sigma m_j - \Sigma \overline{m}_j = 0$. Hence, (23) implies $\Sigma b_{N_j} > 0$ which contradicts the market equilibrium condition $\Sigma b_{N_j} = 0$ Similarly, one can show that $i\overline{\pi} - r < 0$ is inconsistent with market equilibrium The only case which *is* consistent with market equilibrium in the money and bond markets

[7]The value of λ can be made to depend on interest rates and on parameters of the distribution of π without affecting the following results.

is then $i\bar{\pi} - r = 0$, a zero-risk premium on nonindexed bonds.[8]

The fact that $i\bar{\pi} - r = 0$ implies by (21) that $m_j + A_j + ib_{N_j} = 0$ for all j. This has various implications which we explore presently. In the earlier sections we treated V_π as an exogenous parameter and analyzed its effects on the economic system. In the present model, however, p_2 and π are themselves determined by the variability of ΣT_j. Suppose we change the distribution of ΣT_j in such a way that V_π changes while $\bar{\pi}$ remains constant, and analyze the effect of the foregoing change on the individual's optimal behavior, holding all other parameters constant.

As $m_j + A_j + ib_{N_j} = 0$ for all j, we may take this equation combined with the first-order conditions (11) and (12) to determine the individual's m, c, and b_L. Notice that in the quadratic case discussed $f_{\bar{w}} = \alpha - 2\beta\bar{w}$ and none of these equations involve V_π and hence the individual behavior is independent of the variance of the purchasing power of money. This is under the assumption that A_j is not affected by V_π, which seems to be an admissible assumption as it has been observed that in the aggregate $\Sigma A_j = -\Sigma\bar{m}_j$ which depends (directly) only on an economic market variable p_1 and not on the statistical parameter V_π. Suppose for example that $A_j = A$ for all individuals[9] so that $A = -\bar{m}$, where $\bar{m} = \Sigma \bar{M}_j/np_1$. Now, since V_π does not affect individuals' excess demands it cannot affect p_1 and consequently A remains unchanged. Thus under the foregoing assumptions the economic system is invariant to a change in V_π (through ΣT_j), when $\bar{\pi}$ is held constant. In particular an increase in V_π does not create a tendency to shift from nonlinked to linked bonds.

It has been noted (fn. 8) that even under a purely monetary inflation there exist motives to lend and borrow in the nonlinked market. It can be seen however that, under these circumstances, the existence of an active market for nonlinked bonds is based on *differences among individuals* concerning their demands for money or their monetary transfer payments. Thus suppose that $A_j = A = -\bar{m}$. Then by (22), and $i\bar{\pi} - r = 0$, we have $m_j - \bar{m} = -ib_{N_j}$. Thus the variability of b_N among consumers is just the counterpart of the variability of m_j around the population mean. If $m_j = \bar{m}$ for all consumers than $b_{N_j} = 0$ identically and the market for nonlinked bonds disappears.

Another example where homogeneity on the monetary side leads to the disappearance of nonlinked bonds is where the proportion of transfer payments relative to planned money holdings is the same for all individuals, so that $T_j/M_j = \Sigma T_j/\Sigma\bar{M}_j$ for all j. This is equivalent to the government making random "interest payments" (which can be negative) on money holdings. A simple calculation shows that in this case $A_j = -m_j$ so that the condition $m_j + A_j + ib_{N_j} = 0$ implies $b_{N_j} = 0$ for all j. In writing (16) as $w_j = (m_j + t_j)\pi + ib_{N_j}\pi + y_{2j} + rb_{L_j}$ we note that in the present example $(m_j + t_j)\pi$, given m_j, is a constant independent of Π. Thus ("compensated") money ceases to contribute to the riskiness of w. We know however from Section I that this leads to dominance of linked over nonlinked bonds, and hence to the disappearance of the latter market. Thus in order to have a diversified bond market under monetary inflation we need heterogeneity among individuals on the monetary side.

[8]Query: Why should the *lender* ($b_N > 0$) wish to hold nonlinked bonds when they command no premium over the linked ones? The answer lies in the relation between the variability of $b_N\pi$ and $t\pi$. Suppose that for our lender $A_j < 0$ (as it is on the average). Now $Cov(b_{N_j}\pi, t_j\pi) = b_{N_j}Cov(\pi, t_j\pi) = b_N A_j V_\pi < 0$ if $b_{N_j} > 0$ and $A_j < 0$ as assumed. Hence b_N is a *hedge* against variability in the real value of transfer payments $t_j\pi$. It is this hedging property of b_N which constitutes a "premium" for the buyer of nonlinked bonds. It is interesting to note that now b_N is a hedge in two different senses—a negative b_N is a hedge against the variation in the real value of money while a positive b_N is a hedge against the variation in the real value of transfer payments (under the quantity theory assumption).

[9]It can be shown that if $A_j = A$ and $t_j = t$ for all individuals then in equilibrium $V_w = 0$. Thus every individual will find it optimal to manipulate his m_j and t_j so as to eliminate completely the uncertainty in his future wealth.

We should point out that not every kind of endogenous inflation leads necessarily to a zero-risk premium of nonlinked over linked bonds as in the case of a purely monetary inflation. Consider alternatively a "functional inflation" which is intended to finance the transfer of real resources from the private sector to government use. Then we can show that a positive risk premium will prevail in the market. Our reasoning is as follows.

Suppose that in the latter type of inflation the government prints money to purchase goods from the private sector. The variation in the size of these purchases can then be taken as the basic source of random variation of the price level in the economy. However, this government deficit-spending does not represent (in an exchange economy) transfer payments in the sense of *incomes* of individuals but rather payments in exchange of goods. Thus we can express w_j as $w_j = (m_j + ib_{N_j})\pi + rb_{L_j} + (y_{2j} - g_j) + z_j$ where g_j represents the units of the commodity acquired by the government from individual j in exchange for z_j units of real money. Since this is an act of exchange, we have $z_j = g_j$ and the wealth equation becomes again $w_j = (m_j + ib_{N_j})\pi + rb_{L_j} + y_{2j}$.

The mechanism by which the government induced the private sector to give up consumption of real goods is through a reduction in real cash balances represented by $m_j\pi$. This reduction in turn is caused by the inflationary injection of the new money, which equals $(\Sigma z_j)p_2$. (The total supply of real money in the economy in the second period is then given by $\Sigma m_j + (\Sigma z_j)$.)

The previous discussion indicates three components of wealth changes in the second period. The first is the reduction in real endowed commodities; the second is the increase in real money in exchange for the commodities, while the third component is the reduction in the real value of money balances transferred from the first period. The net change equals the last component. It follows from this discussion that the individual wealth equation remains of the same form as in Sections II and III. Consequently, all our previous conclusions concerning the sign of the risk premium and the effect of a change in V_π remain unaltered (in the present model V_π is, of course, a function of the variance of Σg_j).

APPENDIX

Let us derive the expressions for the effect of V_π on our dependent variable in the mixed portfolio (linked, nonlinked) model. The consumer's problem is maximize $U = u(c, m) + \alpha\bar{w} - \beta(\bar{w}^2 + V_w)$ with respect to c, m, and b_L where

$$\bar{w} = (m + ib_N)\bar{\pi} + rb_L + y_2$$
$$V_w = (m + ib_N)^2 V_\pi$$
$$b_N = y_1 - c + \bar{m} - m - b_L$$

The assumptions about the properties of U are stated in the text.

Differentiating U we obtain

$$\text{(A1)} \quad \frac{\partial U}{\partial c} = u_c - ik = 0$$

$$\text{(A2)} \quad \frac{\partial U}{\partial m} = u_m - (i - 1)k = 0$$

$$\text{(A3)} \quad \frac{\partial U}{\partial b_L} = r(\alpha - 2\beta\bar{w}) - ik = 0$$

where $k = (\alpha - 2\beta\bar{w})\bar{\pi} - 2\beta V_m(m + ib_N)$ Assuming a regular maximum we have a negative definite Hessian for the foregoing system. Since the system is of order three the determinant of the Hessian is negative Multiply (A2) by $-i$ and add to it $(i - 1)$ times (A1). Then add minus one times (A3) to (A1). The system is then transformed to

$$\text{(A4)} \quad u_c - r(\alpha - 2\beta\bar{w}) = 0$$
$$\text{(A5)} \quad (i - 1)u_c - iu_m = 0$$
$$\text{(A6)} \quad r(\alpha - 2\beta\bar{w}) - ik = 0$$

The transformed system has a Hessian with a *positive* determinant since in our transformation we had one change of sign (by multiplying (A2) by $-i$).

Differentiating the new system with respect to V_π we obtain (A7), where the coefficients in the third row need not be speci-

(A7)

$$(u_{cc} - 2\beta ri\bar{\pi})\frac{dc}{dV_\pi} + [u_{cm} - (i-1)r2\beta\bar{\pi}]\frac{dm}{dV_\pi} + 2\beta r(r - i\bar{\pi})\frac{db_L}{dV_\pi} = 0$$

$$[(i-1)u_{cc} - iu_{mc}]\frac{dc}{dV_\pi} + [(i-1)u_{cm} - iu_{mm}]\frac{dm}{dV_\pi} + 0\frac{db_L}{dV_\pi} = 0$$

$$a_{31}\frac{dc}{dV_\pi} + a_{32}\frac{dm}{dV_\pi} + a_{33}\frac{db_L}{dV_\pi} = -2\beta i(m + ib_N)$$

fied explicitly. Solving this system by Cramer's Rule, we obtain

$$\frac{db_L}{dV_\pi} = (Det)^{-1}H[-4\beta ri\bar{\pi}(i-1)u_{cm} + 2\beta i^2 ru_{mm} + (i-1)^2 2\beta r\bar{\pi}u_{cc} + i(u_{cm}^2 - u_{cc}u_{mm})]$$

$$\frac{dm}{dV_\pi} = (Det)^{-1}(H[i-1)u_{cc} - iu_{cm} \cdot 2\beta r(r - i\bar{\pi})$$

$$\frac{dc}{dV_\pi} = (Det)^{-1}(-H)[(i-1)u_{cm} - iu_{mm}] \cdot 2\beta r(r - i\bar{\pi})$$

where $H = -2\beta i(m + ib_N) < 0$. Assuming $u_{cm} \geq 0$, strict concavity of $u(c,m)$, $i\bar{\pi} - r > 0$, a positive nominal net interest rate $i - 1 > 0$ and remembering that the determinant of the system is positive, i.e., $(Det) > 0$, we find

$$\frac{db_L}{dV_\pi} > 0, \qquad \frac{dm}{dV_\pi} < 0, \qquad \frac{dc}{dV_\pi} < 0$$

Since $db_N/dV_\pi = -(db_L/dV_\pi + dm/dV_\pi + dc/dV_\pi)$ we obtain from the previous calculations

$$\frac{db_N}{dV_\pi} = -(Det)^{-1}H\{-2\beta r[2i(r - \bar{\pi}) + (i\bar{\pi} - r)]u_{cm} + 2\beta r(i-1)(r - \bar{\pi})u_{cc} + 2\beta ir[r + (i-1)\bar{\pi}]u_{mm} + i(u_{cm}^2 - u_{cc}u_{mm})\}$$

which is negative if $r > \bar{\pi}$.

REFERENCES

G. L. Bach and R. A. Musgrave, "A Stable Purchasing Power Bond," *Amer. Econ. Rev.*, Dec. 1941, *31*, 823–25.

C. G. Fane, "Index Linking and Inflation," *Nat. Inst. Econ. Rev.*, Nov. 1974, No. 70, 42.

S. Fischer, "The Demand for Index Bonds," *J. Polit. Econ.*, June 1975, *83*, 509–34.

M. Friedman, "Using Escalators to Help Fight Inflation," *Fortune*, July 1974, *90*, 94–97.

R. Goode, "A Constant-Purchasing Power Savings Bond," *Nat. Tax J.*, Dec. 1951, *4*, 332–40.

M. Sarnat, "Purchasing Power Risk, Portfolio Analysis, and the Case for Index-Linked Bonds," *J. Money, Credit, Banking*, Aug. 1973, *5*, 836–45.

American Taxation of Multinational Firms

By THOMAS HORST*

Taxation of the multinational firm's foreign income has been debated continually over the last fifteen years. In the early 1950's the U.S. Treasury proposed to eliminate the deferral of *U.S.* taxes on foreign subsidiaries' retained earnings, in order to discourage the flow of direct investment capital abroad and to hasten the repatriation of direct investment income.[1] The Congress was unwilling to take so large a step, but in the Tax Reform Act of 1962 did require dividend income from developed countries to be "grossed up" and did limit the tax-haven abuse of deferral.[2] In the late 1960's the AFL-CIO grew increasingly concerned that *U.S.* multinationals were "exporting jobs" and called for the repeal of deferral and the foreign tax credit.[3] The New Economic Policy announced by President Nixon on August 15, 1971 sought to improve the balance of payments through a variety of measures including a tax preference for export income, the Domestic International Sales Corporation (*DISC*).[4] A primary argument in convincing Congress of the merits of *DISC* was that export income should enjoy a tax deferral comparable to foreign investment income. Although the union-backed Burke-Hartke Bill to repeal deferral and the foreign tax credit was voted down decisively in 1973, the Senate version of the Tax Reduction Act of 1975 would have eliminated deferral. Although this provision was dropped by the House-Senate conference committee, a special subcommittee chaired by Congressman Rostenkowski of the House Ways and Means Committee was established to study deferral and related issues more thoroughly. Finally, the Treasury implemented new guidelines for Sections 861-864 of the Internal Revenue Code which will create a strong tax incentive for *U.S.* investors to charge their foreign subsidiaries more for research and development undertaken by the parent. In opposing all these changes in *U.S.* policy, the multinationals have argued that higher taxes would undermine their competitiveness in world markets without helping *U.S.* exports, employment or the balance of payments.

This paper analyzes the profit-maximizing behavior of a multinational firm and explores the impact on that behavior of repealing deferral, compelling higher charges for *R & D* to foreign subsidiaries, and eliminating the foreign tax credit. My model is limited in obvious respects: the analysis is static, not dynamic; the "multinational" invests at home and in one foreign country; and exports between parent and subsidiary are ignored. Higher *U.S.* taxes on foreign investment income may encourage the firm to invest more at home and less abroad, but that substitution is not necessarily matched by an increase in the parent's ex-

*Fletcher School of Law and Diplomacy. Research was supported by a contract with the U.S. Treasury which endorses neither the methods nor the conclusions of this analysis. Thomas Pugel was instrumental in developing the mathematical analysis and the computer program used in empirical simulations. Gary Hufbauer, James Nunns, and George Kopits have devoted considerable time to earlier versions of the paper and made numerous constructive criticisms.

[1] "Deferral" means taxing the dividends, but not the retained earnings, of a subsidiary. The evolution of *U.S.* taxation of foreign-source income is more fully described in the forthcoming study by C. Fred Bergsten, Thomas Horst, and Theodore Moran, ch. 6. Gary Hufbauer and David Foster provide a thorough analysis of deferral and its relationship to other aspects of current *U.S.* tax policy.

[2] "Grossing up" means basing the tentative *U.S* tax (i.e., before deducting the tax credit) on subsidiaries' dividends inclusive of the foreign income taxes allocable to those dividends.

[3] The foreign tax credit allows *U.S.* investors to reduce their *U.S.* income tax liability by the amount of foreign income and withholding taxes paid to foreign governments; see below.

[4] A *DISC* is essentially a dummy corporation established to receive tax-sheltered export income; half of that income must be paid out as a dividend and thereby subject to *U.S.* taxation, but half may be reinvested in export-related assets, such as export accounts receivable.

orts or a decrease in its imports. The iodel's primary virtue is incorporating everal complex features of *U.S.* tax policy ito a coherent analysis of multinational ivestment behavior. I explore not only the ocation of real investment, but also the ptions for financing that investment. As I iall show, *U.S.* tax policy affects firms' nancial behavior, and that behavior in irn mitigates the impact of tax policy on ie location of new investment. Finally, I ave kept the analysis simple enough to be ble to construct rough estimates of its arameters and simulate the possible imact of various tax changes. In the follow-ig sections I set forth the basic model, escribe the impact of various tax changes, nd summarize the more important conclu-ions. To aid the reader, the mathematical otation is summarized in Appendix Ta-le A1.

I. The Basic Model

It is assumed that a multinational firm tarts with an existing stock of foreign and lomestic investment and seeks the optimal hange in its position over the next year. Because the earnings generated by existing nvestments are an important source of apital for new investment, we cannot ig-iore the role of the past in shaping the resent. To be specific, assume that current enues (net of labor and material costs, gross of interest expenses and income es) R depend on the existing stock of estment I_0, plus new investment under-en during the current period I.

$$R = R(I_0 + I)$$
$$R^* = R^*(I_0^* + I^*)$$

asterisk differentiates the parent's do-stic investment from its foreign sub-iary's. Presumably, the marginal and rage return on investment at home or oad declines as the level of investment ands.

The analysis focuses not only on the els of foreign and domestic investment, t also on the financing of that investment. r simplicity's sake we ignore the market new equity and concentrate on that for debt. Rather than exploring directly the determinants of an optimal debt-equity ratio for the parent, subsidiary, and/or consolidated enterprise, it is merely assumed that new funds are available at home and overseas at increasing rates of interest. These increasing interest rates could reflect either the thinness of local capital markets or lenders' fears of the insolvency of the borrower. In either event, total borrowing costs B consist of those incurred by past borrowing L_0, plus those resulting from new borrowing, L:

$$B = B(L_0 + L) \tag{3}$$
$$B^* = B^*(L_0^* + L^*) \tag{4}$$

The levels of investment and borrowing are linked by balance sheet constraints. For the foreign subsidiary new investment I must equal new borrowing L, plus new funds obtained from the parent F, and the subsidiary's own retained earnings E_R:

$$I = L + F + E_R \tag{5}$$

The parent's new investment I^* must equal its own new borrowing L^*, less new funds advanced to the subsidiary F, plus its own retained earnings:

$$I^* = L^* - F + E_R^* \tag{6}$$

It is also assumed that the foreign subsidiary can deduct from its taxable income interest on *intra*firm debt as well as royalties, headoffice charges and other such payments for intrafirm services. Although I want to postpone for the moment the role of tax avoidance in determining such payments, let me note here that intrafirm interest expenses depend on the intrafirm interest rate i_p, and the ratio of debt to total intrafirm transfer of capital (debt plus equity) f. Total intrafirm interest payments equal those on past borrowing, $i_{p0}f_0F_0$, plus those on new borrowing, i_pfF. Likewise, it will be assumed that royalties, headoffice, and other intrafirm charges vary at least in the short run in proportion h to the foreign subsidiary's total investment, $I_0 + I$. Thus, the foreign subsidiary's taxable income E_B equals the revenues from investment R, minus the interest paid to outside lenders B,

the interest paid to the *U.S.* parent $i_{p0}f_0F_0 + i_pfF$, and payments for royalties, headoffice services, and the like, $h(I_0 + I)$:

$$E_B = R - B - (i_{p0}f_0F_0 + i_pfF) - h(I_0 + I) \tag{7}$$

Next it is assumed that the subsidiary pays income taxes at the rate t, and that the dividends paid to the parent D are some proportion p of income after taxes:

$$D = p(1 - t)E_B \tag{8}$$

Thus, the retained earnings available for reinvestment by the foreign subsidiary E_R, as shown in equation (5) above, are

$$E_R = (1 - p)(1 - t)E_B \tag{9}$$

Most foreign governments collect not just an income tax, but also withholding taxes (a typical rate would be 10 to 15 percent) on dividends, interest, royalties, and other payments to *U.S.* investors. Total withholding taxes paid W, equal:

$$W = w_DD + w_B(i_{p0}f_0F_0 + i_pfF) + w_Hh(I_0 + I) \tag{10}$$

where w_D, w_B, and w_H are the withholding tax rates for dividends, interest, and royalties, headoffice charges, etc., respectively.

We can now turn to *U.S.* tax policy. Rather than taxing foreign investment income net of foreign income and withholding taxes, the United States bases its tax on foreign-source income gross of foreign taxes and then grants a tax credit for foreign taxes paid. In my notation *U.S.* taxable income E_B^* equals domestic income net of interest costs $R^* - B^*$, plus interest, royalties, or other such receipts for intrafirm services and dividends:

$$E_B^* = R^* - B^* + (i_{p0}f_0F_0 + i_pfF) + h(I_0 + I) + D/(1 - t) \tag{11}$$

Note that intrafirm income receipts are *not* reduced by the withholding tax and that dividends have been grossed up to include the foreign income taxes "deemed paid" on those dividends.[5] Foreign-source income for the purposes of determining *U.S.* taxes exceeds the cash actually received by the parent. Note further that *U.S.* taxable income includes the dividends, but not the retained earnings of the foreign subsidiary. That is the essence of deferral.

The taxes paid to the U.S. Treasury T^*, equal the *U.S.* tax rate t^* times taxable income E_B^*, less the foreign tax credit T_C^*:

$$T^* = t^*E_B^* - T_C^* \tag{12}$$

The *U.S.* investor can claim a foreign tax credit equal to the *lesser* of two amounts: 1) the withholding taxes on dividends, interest, royalties, etc. plus the income taxes deemed paid on the dividends; and 2) the *U.S.* tax rate[6] times total foreign-source income:

$$T_C^* = min\left\{W + \frac{tD}{1 - t}, t^* \cdot [i_{p0}f_0F_0 + i_pfF + h(I_0 + I) + D/(1 - t)]\right\} \tag{13}$$

If foreign taxes paid or deemed paid are less than the maximum creditable—the *U.S* taxes which would have been due on the foreign-source income—the investor is said to have a deficit of tax credits. Using the withholding tax formula (10), we can prove that the investor will have a deficit of foreign tax credits if and only if the share-weighted average of foreign tax rates is less than the *U.S* income tax rate:

$$s_Bw_B + s_Hw_H + s_D[w_D(1 - t) + t] < t^* \tag{14}$$

where s_B, s_H, and s_D sum to unity and are the shares of interest, royalties, and other such fees, and dividends, respectively, in the *U.S.* investor's foreign-source income. Condition (14) indicates that even if the foreign tax burden on dividend income $w_D(1 - t) + t$ exceeds the *U.S.* tax rate t^*, the investor may yet avoid having surplus foreign tax credits by making sufficiently large interest and royalty payments. Let us rewrite the formula for the foreign tax credit as:

[5]Legally speaking, the foreign subsidiary pays the income tax, and the *U.S.* parent is only deemed to have paid those taxes.

[6]Actually the foreign tax credit is limited not by the statutory tax rate but by the ratio of the firm's tentative *U.S.* taxes (i.e., before tax credits) to its total taxable income. If the firm has capital gains or certain other favorably taxed income, the statutory rate may exceed its average tax rate.

$$(15)\quad T_C^* = W + \frac{tD}{1-t} - x\{(w_B - t^*)(i_{p0}f_0F_0 + i_pfF) + (w_H - t^*)h(I_0 + I) + [w_D(1-t) + t - t^*]D/(1-t)\}$$

where x is a binary variable equal to zero if and only if the investor has a deficit of foreign tax credits, i.e., condition (14) is satisfied. While equation (15) looks messier than (13), it is more tractable analytically.

The *U.S.* parent's after-tax income E_A^* equals its before-tax income E_B^*, less its *U.S.* taxes T^*, and the foreign taxes included in its taxable income:

$$(16)\quad E_A^* = E_B^* - T^* - [W + tD/(1-t)]$$

The consolidated, after-tax income of the multinational enterprise equals the parent's after-tax income (which includes dividend income from its foreign subsidiary) plus the subsidiary's retained earnings:

$$(17)\quad E_C^* = E_A^* + E_R$$

To close out the model, we need to specify the parent's dividends and consequently its earnings available for reinvestment. It is assumed that the dividends paid to the ultimate shareholders D^* are some constant proportion p^* of *consolidated* after-tax earnings:

$$(18)\quad D^* = p^*E_C^*$$

The parent's retained earnings are the difference between its after-tax income and its dividends:

$$(19)\quad E_R^* = E_A^* - D^* = (E_C^* - E_R) - p^*E_C^* = (1-p^*)E_C^* - E_R$$

From (19) it is apparent that consolidated after-tax earnings E_C^* are proportionate to consolidated retained earnings available for investment:

$$(20)\quad E_C^* = \frac{1}{1-p^*}(E_R^* + E_R)$$

It is assumed that the multinational firm seeks to maximize either consolidated after-tax earnings or, equivalently, consolidated retained earnings available for investment. I have little direct evidence on what, if anything, a multinational strives to maximize and have chosen consolidated after-tax earnings E_C^* because it seems to be as reasonable and as convenient an objective as any. I should note, however, that this objective function does *not* discount the value of earnings retained abroad despite the probable tax cost of repatriating those funds as dividends. Thus the behavioral assumption would most aptly characterize a management-controlled firm whose primary objective was the growth of the firm and for whom dividends to shareholders are comparable to a tax on consolidated earnings.

What should the firm do to maximize its consolidated earnings? To begin to answer this question, let us substitute several of the earlier formulas into equations (17) and (20) and rewrite consolidated after-tax earnings as:

$$(21)\quad E_C^* = \frac{E_R^* + E_R}{1-p^*}$$
$$= (1-t^*)(R^* - B^*) + (1-t)(R-B) - [t^* - t + x(w_B - t^*)](i_{p0}f_0F_0 + i_pfF) - [t^* - t + x(w_H - t^*)]h(I_0 + I) - [t^* - t + x(w_D(1-t) + t - t^*)]\,pE_R/(1-t)(1-p)$$

The multinational firm has *seven* degrees of freedom in maximizing consolidated income. It can set four intrafirm financial parameters—i_p, f, h, and p—plus the rates of new domestic and foreign investment I^* and I, and the rate of transferring new capital to its subsidiary F. All other variables in equation (21) are either predetermined or will be determined by these seven values.

We distinguish the four intrafirm financial parameters, i_p, f, h, and p, from the three remaining controls, I^*, I, and F, because the former should be minimized or maximized according to straightforward, tax-avoidance criteria, while the latter must satisfy standard first-order conditions. Let us look first at the intrafirm financial parameters for their optimal values can be determined by a close inspection of equation (21). We obviously need to differentiate between cases where an investor has a defi-

cit, rather than a surplus, of foreign tax credits.

Deficit Case: A deficit of tax credits obtains when foreign tax rates are comparatively low, condition (14) can be satisfied, and thus $x = 0$. In this case the firm should *minimize* royalties, interest, dividends, and all other forms of repatriating income. Foreign investment should be financed insofar as possible out of retained earnings rather than new funds obtained from the parent. If funds must be repatriated in one form or another, the investor will, so far as taxes are concerned, be indifferent among interest, royalties, headoffice charges, or dividends. Each generates the same increase in total tax payments.

Surplus Case: A surplus of tax credits obtains when foreign tax rates are comparatively high, condition (14) is violated, and thus $x = 1$. In this case the firm should *maximize* royalties, headoffice charges, interest payments, and all other intrafirm charges deductible from the foreign subsidiary's income and *minimize* dividend payments. By substituting one form of income repatriation for another, the investor can reduce its foreign income taxes without a corresponding increase in its *U.S.* taxes. A multinational firm, in short, has a clear tax incentive to manipulate its intrafirm accounts to avoid generating excess tax credits.

Let me hasten to add that multinational firms have less than full flexibility in manipulating intrafirm accounts and that in actual practice tax avoidance is not the only criterion affecting the firm's behavior. National tax authorities strive to protect the local tax base: foreign governments may limit deductible payments to parent firms, and the United States applies its "arm's-length" standard to intrafirm interest rates, royalties, etc..[7] Likewise, foreign exchange authorities frequently limit intrafirm transactions to improve the balance of payments —witness the *U.S.* balance-of-payments guidelines affecting dividend repatriation in the late 1960's. Furthermore, the multinational may be willing to pay higher taxes in order to withdraw its income from weak-currency countries or to minimize its exposure to expropriation. But taxes do matter. Sidney Robbins and Robert Stobaugh (pp. 28-29 and 77) found that intrafirm debt-equity ratios and the methods of income repatriation reflected tax considerations. George Kopits (1972, 1974) found that subsidiaries' dividend payout rates were higher the lower the tax cost of paying dividends and that royalty rates were manipulated to offset excess tax credits generated by dividends.

Consolidated income depends also on the rates of new domestic and foreign investment I^* and I, and new funds advanced to the foreign subsidiary F. The first-order condition obtained by taking the partial derivative of E_C^* with respect to domestic investment I^* is:[8]

$$r^* = b^* \tag{22}$$

The marginal revenue from new domestic investment r^* should equal the marginal cost of newly borrowed funds in the United States, b^*. The analogous condition with respect to new foreign investment I is slightly more complicated:

$$r = b + \frac{(t^* - t_f)h}{(1 - t_f)} \tag{23}$$

where t_f is defined to be the effective rate of global (foreign plus *U.S.*) taxation of the foreign subsidiary's income. Assuming the investor succeeds in avoiding a surplus of foreign tax credits, this effective rate of taxation is simply a weighted average of the foreign and *U.S.* income tax rates:

$$t_f = pt^* + (1 - p)t = t + p(t^* - t) \tag{24}$$

The portion of foreign subsidiary earnings paid out as dividends p, is taxed at the *U.S.*

[7]The arm's length standard obligates an American investor to use transfer prices equal to those which would have prevailed between an independent buyer and seller.

[8]The derivations of the first-order conditions and the equations of change are straightforward, but exceedingly cumbersome. Because they will be published in the Mathematical Appendix to Bergsten, Horst, and Moran, ch. 6, and are available from the author on request, they are not reproduced here.

rate t^*, while that portion retained by the subsidiary $1 - p$, is taxed at the foreign rate t. The second term on the right-hand side of condition (23) recognizes that increased foreign investment generates higher headoffice and other such charges and that such income will be taxed at the *U.S.*, rather than the foreign, rate. Practically speaking, this second term on the right-hand side of condition (23) is small, and the foreign subsidiary equates the marginal revenue from new investment to the marginal cost of locally borrowed funds.

The third and most interesting first-order condition is obtained from the derivative of consolidated income with respect to new funds advanced to the subsidiary F:

$$(25) \qquad (1 - t^*)b^* = (1 - t_f)b - (t^* - t_f)i_p f$$

The significance of this condition is grasped more readily if we contrast two extreme cases, the equity-only and the debt-only investor. Suppose, first, that an American company invests overseas, borrows in local capital markets, but limits its own investment in its subsidiary to equity participation. That is to say, there is no intrafirm debt, only equity. This method of financing foreign investment is certainly encouraged by current *U.S.* tax policy, since debt gives rise to interest payments and thereby shifts income to the more highly taxed *U.S.* parent. In this equity-only case, $f = 0$, and condition (25) reduces to:

$$(26) \qquad (1 - t^*)b^* = (1 - t_f)b$$

The optimal transfer of new funds to the subsidiary equates the after-tax return on equity in the two countries. If the *U.S.* income tax rate t^* exceeds the foreign income tax rate t, it will also exceed t_f, and the foreign investor will have an implicit tax incentive to invest abroad rather than in the United States.

Suppose for the sake of contrast that the American investor disregards the tax advantages of equity and relies exclusively on debt to finance new foreign investment. Suppose further that it was willing and able to charge interest equal to the subsidiary's marginal cost of newly borrowed funds b. In this case, $f = 1$, $i_p = b$, and condition (25) reduces to:

$$(27) \qquad b^* = b$$

The multinational has no tax incentive to favor foreign investment over domestic, for *at the margin* they both generate equal tax payments. While deferral may reduce *U.S.* taxes and thus leave the multinational with additional funds for global investment, it would not bias the location of that investment. Although condition (27) represents a hypothetical extreme, it points up an important conclusion: deferral encourages an American investor to favor foreign over domestic investment 1) the lower the foreign income tax rate, 2) the lower the rate of dividend repatriation, 3) the lower the ratio of debt to new capital (debt plus equity) transferred to the subsidiary, and 4) the lower the interest charged on intrafirm debt. Deferral thus encourages the American investor to use equity rather than debt in financing foreign investment, and the more equity used, the greater is the implicit bias towards foreign and against domestic investment.

II. The Impact of Changing U.S. Tax Policy

A. *Repealing Deferral*

In this section we will explore the impact of three proposed changes in current *U.S.* tax policy: repealing deferral; increasing *R&D* charges to foreign subsidiaries; and eliminating the foreign tax credit. If deferral were repealed, *U.S.* taxable income would include the grossed-up value of the foreign subsidiary's retained earnings $E_R/(1 - t)$ in addition to the income shown in equation (11) above. The *U.S.* taxable income and consolidated before-tax income now become one and the same:

$$(28) \qquad E_B^* = (R^* - B^*) + (R - B)$$

Assuming the foreign tax credit would be extended to include all foreign income taxes paid, not just those associated with dividends, consolidated after-tax income would equal:

$$(29)\quad E_C^* = (1 - t^*)(R^* - B^* + R - B) - x\Big\{(w_B - t^*)(i_{p0} f_0 F_0 + i_p f F) + (w_H - t^*)h(I_0 + I) + [w_D p(1 - t) + t - t^*]\frac{E_R}{(1 - p)(1 - t)}\Big\}$$

As long as foreign income and withholding tax rates are low enough, or royalties, interest, and other such charges can be kept high enough to avoid a surplus of tax credits, $x = 0$, and the multinational firm has no further tax incentives to minimize or maximize the use of debt, intrafirm interest rates, royalties, dividends, or the like. The only tax consideration affecting the intrafirm accounts would be the desire to avoid excess tax credits.

Turning from the intrafirm financial parameters to the rates of new domestic and foreign investment and new capital transferred to the subsidiary, the first-order conditions for maximizing consolidated income are:

$$(30)\quad r^* = b^*$$

$$(31)\quad r = b$$

$$(32)\quad b^* = b$$

All the tax terms have cancelled each other out. Without deferral marginal revenue from new investment should equal the marginal cost of newly borrowed funds both within and between the two countries. Gone would be the implicit tax incentive to invest abroad rather than in the United States.[9]

To determine the amounts by which foreign and domestic investment might change, we must take the total derivatives of the original first-order conditions, (22), (23), and (25). Assuming for the moment that the intrafirm financial ratios i_p, f, h, and p remain fixed, we obtain three equations in three unknowns (dI^*, dI, and dF) and one known $dt_f = (1 - p)(t^* - t)$. Rather than presenting the messy formulas,[10] let me describe the nature of the calculations and then proceed to some numerical examples. The impact of eliminating deferral can be decomposed into a *substitution* and *liquidity* effect. When *U.S.* taxes due on foreign investment income increase, the investor diverts fewer funds to foreign investment and more to domestic. This is the substitution effect. The liquidity effect derives from higher taxes: with fewer funds available for reinvestment and with outside funds becoming increasingly costly, the multinational cuts back on its global investment. The substitution and liquidity effects both reduce the rates of new foreign investment and new funds transferred to the foreign subsidiary. New domestic investment will increase if the positive substitution effect outweighs the negative liquidity effect.

My numerical examples are based upon the 1974 experience of American-owned manufacturing subsidiaries—see Appendix Table A2. Most of our model's parameters were roughly estimated from published statistical sources. For example, the average foreign income tax of 39 percent and average dividend payout rate of 42 percent implied an effective rate of global taxation of foreign subsidiary income of 42.8 percent. Thus, eliminating deferral would raise the effective rate of taxation from 42.8 percent to the *U.S.* rate of 48 percent, an increase of 5.2 percent. The parameters I could *not* estimate from any published sources were the elasticities of investment demand or borrowing costs implicit in equations (1) through (4) above. Accordingly, I arbitrarily assumed that the elasticities of the marginal revenue from new investment and marginal cost of new borrowing were equal to two and then compared those results with others based on higher or lower elasticities.

The numerical calculations of the impact of eliminating deferral are summarized in Table 1. Notice that I have scaled my example to make it appear that all new investment by foreign manufacturing affiliates of

[9]In actual practice, various nonneutralities would remain. For example, the *U.S* investment tax credit does not apply to foreign investment, and state and local income taxes are only deductible from *U.S.* federal income taxes, while analogous foreign taxes would be creditable. These remaining nonneutralities are carefully evaluated by Hufbauer and Foster.

[10]See fn. 8 above.

TABLE 1—ESTIMATED IMPACT OF REPEALING DEFERRAL: ON NEW DOMESTIC AND FOREIGN INVESTMENT, NEW FUNDS ADVANCED TO SUBSIDIARIES, CONSOLIDATED AFTER-TAX INCOME, AND DOMESTIC AND FOREIGN TAXES PAID BY U.S. MANUFACTURERS, 1974

		Case 1		Case 2 (More Elastic Investment)		Case 3 (Less Elastic Borrowing)	
	Initial Value	Absolute Change	Percentage Change	Absolute Change	Percentage Change	Absolute Change	Percentage Change
Domestic Investment	36,400	1,429	3.9	3,613	9.9	1,360	3.7
Foreign Investment	18,300	-1,549	-8.5	-3,789	20.7	-1,579	-8.6
New Funds for Subsidiary	2,710	-2,466	-91.0	-4,642	-171	-2,028	-75
onsolidated After-Tax Income	15,194	-532	-3.5	-477	-3.1	-534	-3.5
S Taxes Paid	6,005	545	9.1	501	8.3	532	8.9
oreign Taxes Paid	5,001	-80	-1.6	-148	-3.0	-65	-1.3

te Initial Values and Absolute Changes expressed in millions of dollars. Case 1 assumes the values for param-rs shown in Appendix Table A2. Case 2 assumes the same values except the values of $r^{*\prime}$ and r' are two-fifths large as those shown in Appendix Table A2. Case 3 assumes the same values as shown in Table A2 except the lues of $b^{*\prime}$ and b' are twice as large as those shown.

U.S. corporations in 1974 was undertaken by the sole subsidiary of a large U.S. manufacturer. In Case 1 (where all investment and borrowing elasticities are assumed to equal two) eliminating deferral would increase the parent's investment by 3.9 percent and reduced the subsidiary's by 8.5 percent in 1974. Because of lost liquidity the rate of global investment would have been slightly lower than it actually was. The substantial reduction in intrafirm funds transferred reflects our implicit assumption that eliminating deferral would encourage multinational firms to borrow more abroad and less at home than they now do. That is to say, taxes affect the location of borrowing as well as the location of investment. Consolidated after-tax income would have been $532 million less than it was, which would represent a 3.5 percent decline for the large multinational manufacturer. These lowered earnings are the by-product, of course, of the higher taxes paid to the U.S. government. Foreign tax payments ease slightly because of the lower rate of new foreign investment and the higher rate of new borrowing by the foreign subsidiary.

In Case 2 all the parameters are identical to Case 1 except that we have assumed that new domestic and foreign *investment* are *more elastic* with respect to changes in the cost of capital than they were in Case 1 (the elasticities are equal to five rather than two). As can be seen, the more elastic domestic and foreign investment are, the greater is the substitution of domestic for foreign investment resulting from the loss of deferral. Case 3 differs from Case 1 only in assuming that new domestic and foreign *borrowing* are *less elastic* with respect to changes in the interest rate. As one can see, foreign investment falls by more and domestic rises by less than they did in Case 1. In short, the more elastic investment demand the greater the substitution effect is, and the more elastic the supply of external funds the smaller the liquidity effect is.

The estimates in Table 1 all assume that the intrafirm financial ratios i_p, f, h, and p remain fixed at their current values. As we noted above, however, eliminating deferral might also encourage investors to rely more on debt and less on equity in financing their foreign investment, to raise intrafirm interest and royalty rates, to expand charges for *R&D* and other headoffice expenses and to increase dividend payout rates. In the absence of deferral, all these changes except higher dividends would shift tax revenues from the foreign country to the United States. The statistics in Table 1 could thus understate the impact of eliminating de-

TABLE 2—ESTIMATED IMPACT OF DOUBLING HEADOFFICE, *R&D*, AND OTHER INTRAFIRM SERVICE CHARGES: ON NEW DOMESTIC AND FOREIGN INVESTMENT, NEW FUNDS ADVANCED TO SUBSIDIARIES, CONSOLIDATED AFTER-TAX INCOME, AND DOMESTIC AND FOREIGN TAXES PAID BY *U.S.* MANUFACTURERS, 1974

		Case 1 (Deductions Allowed)		Case 2 (Deductions Disallowed)	
	Initial Value	Absolute Change	Percentage Change	Absolute Change	Percentage Change
Domestic Investment	36,400	149	0.4	1,393	3.8
Foreign Investment	18,300	−332	−1.8	−3,087	−16.9
New Funds for Subsidiary	2,710	444	16.4	−2,718	−100
Consolidated After-tax Income	15,194	−142	−0.9	−991	−6.5
U.S. Taxes Paid	6,005	688	11.5	981	16.3
Foreign Taxes Paid	5,001	−592	−11.8	−84	−1.7

Note: Initial Values and Absolute Changes expressed in millions of dollars. Both cases assume that headoffice and other such charges are raised from 1.1 to 2.2 percent of the foreign subsidiaries' total assets. In Case 1 the foreign government allows higher deductions from the subsidiaries' taxable income; in Case 2 they do not.

ferral on both domestic and foreign tax revenues.[11]

B. *Increasing R&D Charges to Foreign Subsidiaries*

As noted in the introduction, the Treasury has issued new guidelines for Sections 861–864 of the Internal Revenue Code. These new guidelines will require the multinationals to allocate a higher portion of their domestic *R&D* expenses to their foreign affiliates when determining their maximum allowable tax credit. Unless the investor has a deficit of tax credits, its *U.S.* tax payments will rise. The Treasury hopes that the multinationals would increase their *R&D* charges to avoid double taxation of this portion of their income. The multinationals claim that foreign tax authorities will not allow any additional deductions from the subsidiaries' income, so the disputed expenses would give rise to double taxation. We have used our microeconomic model to determine the impact of increasing *R&D* or other intrafirm charges assuming first that the foreign government would, and second that it would not, permit higher deductions from foreign subsidiary income.

Our findings are summarized in Table 2. Because there is no way of knowing how great the increase in intrafirm charges might be, we have arbitrarily assumed that all charges except for interest would be doubled. That is to say, headoffice, royalties and all other intrafirm charges would increase from their current 1.1 percent to 2.2 percent of foreign subsidiaries' assets. Case 1 in Table 2 assumes that the foreign government would allow the multinational to deduct all such charges from the foreign subsidiaries' taxable income. As can be seen, the estimated impact on domestic and foreign investment would be minimal. The primary consequence would be a shift of taxable income from the foreign subsidiary to the *U.S.* parent. Foreign tax payments would fall by 12 percent while *U.S.* tax payments would increase by 11 percent. Because the *U.S.* income tax rate is higher than the foreign income tax rate, the global tax burden would rise and consolidated after-tax income would fall slightly.

This impact contrasts sharply with that in Case 2 where we have assumed that the foreign government would *not* allow increased deductions from the subsidiary's taxable income. The key to understanding this latter situation is recognizing that the new guidelines would subject foreign investment in-

[11]Of course, factors which we have ignored could reverse this conclusion. For example, foreign governments might retaliate by raising their taxes on *U.S.*-owned subsidiaries' income, which would increase their parents' *U.S.* tax credits and decrease their tax payments

ome to a disguised form of double taxaon. Foreign investment would be cut by 17 ercent, or twice the reduction from elimiating deferral. Likewise, *U.S.* taxes would icrease by 16 percent, a gain based on the irect impact of the new guidelines and the iduced cutback in funds advanced to the ubsidiary (if more investable funds are reined by the parent, *U.S.* borrowing and iterest expenses can be cut proportiontely). But in this second case *U.S.* tax reveues gain at the expense of the American inestor and not the foreign treasury. In hort, by disallowing higher deductions for R&D expenses, the foreign government can rotect its tax base. But in doing so, it permits the double taxation which inhibits new nvestment by American-owned subsidaries.

C. *Repealing the Foreign Tax Credit*

Several critics of *U.S.* tax policy have roposed the repeal of the foreign tax credit as well as deferral. Taxable income in the United States would include all foreign investment income net of foreign income and withholding taxes:

$$E_B^* = (R^* - B^*) + (R - B - T - W) \tag{33}$$

U.S. taxes would equal the *U.S.* tax rate times the taxable income with no foreign tax credit:

$$T^* = t^* E_B^* \tag{34}$$

Consolidated after-tax income would equal:

$$\begin{aligned} E_C^* = {} & (1 - t^*)(R^* - B^*) \\ & + (1 - t^*)(1 - w_D p)(1 - t)(R - B) \\ & + (1 - t^*)(t + w_D p(1 - t) - w_B) \cdot \\ & (i_{p0} f_0 F_0 + i_p f F) \\ & + (1 - t^*)(t + w_D p(1 - t) - w_H)h(I_0 + I) \end{aligned} \tag{35}$$

Note the heavy taxation, $(1 - t^*)(1 - w_D p) \cdot (1 - t)$, of foreign investment income $R - B$. As is apparent from a close inspection of equation (35), the American investor can raise its consolidated after-tax earnings by increasing its use of intrafirm debt f, the rate of interest on that debt, i_p, or the rate of charging headoffice and other such expenses back to its subsidiary. Each of these measures shifts income from the subsidiary to the parent and relieves the double taxation of foreign investment income.

The new first-order conditions for maximizing after-tax income are:

$$r^* = b^* \tag{36}$$

$$r = b - \frac{t + w_D p(1 - t) - w_H}{(1 - t)(1 - w_D p)} h \tag{37}$$

$$\begin{aligned} b^* = {} & (1 - t)(1 - w_D p)b \\ & + (t + w_D p(1 - t) - w_B) i_p f \end{aligned} \tag{38}$$

Once again, the most interesting of these first-order conditions is the third, and its significance becomes more apparent by contrasting an equity-only with a debt-only investor. If an American investor ignored the obvious tax incentives to finance foreign investment with debt and advanced only equity funds to its subsidiary, $f = 0$ and equation (38) becomes:

$$b^* = (1 - t)(1 - w_D p) b \tag{39}$$

This is the mathematical formula for the AFL-CIO's dream and the multinationals' nightmare! Because of the double taxation of foreign equity income, the American manufacturer discriminates heavily against foreign investment and in favor of *U.S.* investment. By contrast, if an investor could rely wholly on debt in financing new investment and charge interest equal to its subsidiary's marginal cost of borrowing, $i_p f = b$ and equation (38) reduces to:

$$b^* = (1 - w_B)b \tag{40}$$

Although the American investor would still discriminate against foreign investment, that discrimination would be much less than above. The firm's ability to substitute debt for equity in financing foreign investment will limit the substitution of domestic for foreign investment.

Our numerical examples bring out this point quite clearly. The estimates presented in Table 3 show the impact of repealing deferral and the foreign tax credit and allowing only a deduction for foreign taxes. In Case 1 the firm continues its traditional mix of debt and equity in financing foreign expansion; in Case 2 it uses only debt in fi-

TABLE 3—ESTIMATED IMPACT OF REPEALING DEFERRAL AND THE FOREIGN TAX CREDIT AND ALLOWING ONLY A DEDUCTION FOR FOREIGN TAXES PAID: ON NEW DOMESTIC AND FOREIGN INVESTMENT, NEW FUNDS ADVANCED TO SUBSIDIARIES, CONSOLIDATED AFTER-TAX INCOME, AND DOMESTIC AND FOREIGN TAXES PAID BY *U.S.* MANUFACTURERS, 1974

		Case 1 (Initial Parameters)		Case 2 (Reliance on Debt)	
	Initial Value	Absolute Change	Percentage Change	Absolute Change	Percentage Change
Domestic Investment	36,400	9,291	25.5	3,970	10.9
Foreign Investment	18,300	−10,283	−56.2	−4,997	−27.3
New Funds for Subsidiary	2,710	−15,725	−580.3	−8,060	−297.4
Consolidated After-Tax Income	15,149	−2,974	−19.6	−3,107	−20.5
U.S. Taxes Paid	6,005	3,028	50.4	2,953	49.2
Foreign Taxes Paid	5,001	−504	−10.1	−144	−2.9

Note. Initial Values and Absolute Changes expressed in millions of dollars. Estimates include the repeal of deferral—see first three columns of Table 1. Both cases assume that the foreign tax credit is replaced by a deduction from taxable income. In Case 1 all parameters are as shown in Appendix Table A2; in Case 2 the investor raises the ratio of new debt to new funds f to unity, and the interest on that debt to equal the subsidiary's marginal cost of borrowing.

nancing new foreign investment and charges interest equal to the subsidiary's marginal cost of borrowing. The substitution of domestic for foreign investment is far greater in Case 1 (where the firm relies on its traditional debt-equity mix) than in Case 2 (where it shifts from equity to debt). Notice the substantial increase in *U.S.* taxes and the consequent decrease in consolidated corporate earnings and new investment. In fact, had the elasticities of investment demand and loanable funds been less than we supposed, the lost liquidity could more than offset the substitution effect, and domestic investment would *fall* with the repeal of the foreign tax credit. Finally, we note that the intrafirm flow of funds, already diminished by the repeal of deferral, would reverse its direction. The subsidiary either advances funds to the *U.S.* parent or repatriates previously received capital. In actual practice, this backflow of foreign direct investment might be blocked and the substitution of domestic for foreign investment diminished.[12]

[12]The foreign tax authorities, for example, might treat the repatriation of capital as dividends paid out of past earnings and thus subject to a withholding tax. Likewise, foreign subsidiaries' access to local capital markets might be curtailed if local borrowing exceeded local investment.

III. Conclusion

We have explored the effects of *U.S* taxation of foreign investment income on foreign and domestic investment, the financing of that investment, the mix of debt and equity in advancing funds to a foreign subsidiary, and dividends, royalties, and other such intrafirm payments. Under existing *U.S.* policy, corporate income taxes on foreign income are deferred until the subsidiary formally pays a dividend to its *U.S* parent, and the parent can claim a tax credit for income and withholding taxes paid to the foreign government. This policy encourages the use of equity rather than debt in financing foreign expansion and discourages the repatriation of foreign investment income; if income is repatriated, investors may want to make minimal interest, royalty, and other such payments to offset the excess tax credits generated by dividends. Current policy also offers an implicit tax incentive to investing overseas, and the size of the incentive increases the lower the foreign income tax, the lower the subsidiary's dividend payout rate, the less debt and the more equity used in financing foreign investment, and the lower the interest on intrafirm debt.

If deferral were repealed, so too would most of the tax incentives to invest abroad

ır to manipulate intrafirm accounts to void taxes. The impact of this and other ax changes can be disaggregated into a substitution and a liquidity effect. If the elasticities of investment opportunities at home ind abroad are low, so too will be the induced substitution of domestic for foreign nvestment. Likewise, if the elasticities of he supply of outside funds are small, any ax increase will exercise a larger drag on he firm's rate of global investment.

We also explored the effects of compelling *U.S.* investors to charge their foreign ubsidiaries with a higher proportion of current *R&D* expenses and of repealing the foreign tax credit. If foreign governments will allow higher deductions from the subsidiaries' taxable income, the United States would gain tax revenues at the expense of e foreign government. If higher deduc-ons are disallowed, foreign investment in-ome is subject to a disguised form of doue taxation, *U.S.* tax payments rise, rporate after-tax income falls, and the te of foreign investment contracts. Re-aling the foreign tax credit would have a rofound impact on a multinational's profability and, if its ability to compensate for e lost liquidity with increased borrowing small, on its rate of global investment. he substitution of domestic for foreign in-estment depends not only on the elastici-es of investment opportunities, but also e firm's ability to substitute debt for eqity in its internal financing of foreign in-estment.

This analysis can be extended in several irections. Additional foreign countries ould be included; intrafirm exports might e taken into explicit account; other aspects f *U.S.* and foreign tax policy (for example, lepreciation allowances, investment tax redits) could be explored; and better numerical estimates of the model's parameters could be developed. The analysis should incorporate the dynamics of investment planning even though the mathematics could become formidable. Finally, our characterization of a multinational's financial behavior was rudimentary. But the analysis showed that tax policy affects financial behavior, and financial behavior in turn mitigates the impact of tax policy on investment spending. This interaction between tax policy and multinational finance surely deserves further attention.

REFERENCES

C. Fred Bergsten, Thomas Horst, and Theodore Moran, *American Multinationals and American Interests*, Washington, forthcoming.

G. C. Hufbauer and D. Foster, "U.S. Taxation of the Undistributed Income of Controlled Foreign Corporations," Office of International Tax Affairs, Dept. of Treasury, Apr. 1976.

G. F. Kopits, "Dividend Remittance Behavior Within the International Firm: A Cross Country Analysis," *Rev. Econ. Statist.*, Aug. 1972, *54*, 339–42.

——, "Intrafirm Royalties Crossing Frontiers and Transfer Pricing Behavior," unpub. manuscript, Nov. 1974.

M. E. Kyrouz, "Foreign Tax Rates and Tax Bases," *Nat. Tax. J.*, Mar. 1975, *28*, 61–80.

J. R. Nunns and G. C. Hufbauer, "The United States Balance of Tax Payments on Foreign Investment and Employment," *Colum. J. World Bus.*, Summer 1975, *10*, 12–20.

Sidney M. Robbins and Robert V. Stobaugh, *Money in the Multinational Enterprise*, New York 1973.

U.S. Congress, Senate, Committee on Finance, *Implications of Multinational Firms for World Trade and Investment and for United States Trade and Labor*, Washington 1973.

U.S. Office of Business Economics, "U.S. Direct Investment Abroad in 1974," *Surv. Curr. Bus.*, Oct. 1974, *55*, 43–63.

——, "Capital Expenditures by Majority-Owned Foreign Affiliates of U.S. Companies: 1975 and 1976 and 1966–76 Trends," *Surv. Curr. Bus.*, Mar. 1976, *56*, 20–29.

U.S. Office of Foreign Direct Investment, *Foreign Affiliate Financial Survey*, Washington 1971.

APPENDIX

TABLE A1—SUMMARY OF MATHEMATICAL NOTATION

Symbol	Meaning
R	Total return on investment net of labor and material costs, but gross of interest expense and income taxes
I	Investment
B	Total interest paid on external debt
L	Borrowing in external capital market
F	Transfer of investable funds (debt plus equity) from parent to subsidiary
E_R	Retained earnings available for new investment
i_p	Interest rate applied to intrafirm debt
f	Proportion of F which is debt rather than equity
h	Ratio of $R\&D$ and other headoffice charges to total foreign investment $I_0 + I$
E_B	Taxable income
T	Income taxes paid
t	Income tax rate
D	Dividends paid
p	Ratio of dividends paid to earnings after income taxes
E_A	Earnings after income taxes
W	Withholding taxes paid
w_D, w_B, w_H	Withholding tax rates applied to intrafirm dividends, interest, and headoffice charges, respectively
T_C	Foreign tax credit allowed
s_D, s_B, s_H	Relative shares of dividends, interest and headoffice charges in total foreign-source income
x	Binary variable equal to unity if and only if the investor has paid more foreign income and withholding taxes than are creditable
E_C^*	Consolidated (i.e., parent plus subsidiary) after-tax income
t_f	Effective global (i.e., foreign plus *U.S.*) rate of taxation of foreign-subsidiary income
r	Marginal revenue from additional investment, i.e., dR/dI
b	Marginal cost of additional outside borrowing, i.e., dB/dL

Notes: The * denotes the parent's *U.S.* operations, the lack of one the subsidiary's foreign operations. The 0 subscript indicates a predetermined stock, the lack of one the current flow.

TABLE A2—VALUES AND SOURCES OF PARAMETER ESTIMATES USED IN SIMULATIONS

Parameter	Value	Definition and Source
b, b^*	.09	Assumed marginal cost of externally borrowed funds for parent or affiliate
p	.42	Subsidiary's dividend payout ratio. According to the U.S. Office of Business Economics (1975), Table 4, p. 51, manufacturing affiliates had a gross dividends-earnings ratio of .40 in 1974. I have raised this estimate marginally to approximate a four- or five-year moving average.
p^*	.33	Parent's dividend payout ratio. This is the ratio of all *U.S.* manufacturing firms' dividends to after-tax earnings. See U.S. Office of Business Economics (1975), p. S-20.
t	.391	Foreign income tax rate. This estimate is based on Kyrouz's realized tax rates, see Bergsten, Horst, and Moran, Table 6-2, col. 1.
t^*	.48	*U.S.* statutory income tax rate.
E_B	10.67	Affiliate's before-tax earnings. According to the U.S. Office of Business Economics (1975), Table 11, p. 51, the after-tax earnings of manufacturing affiliates in 1974 was $6.498 billion. Since this is net of income taxes, I have grossed this figure up by dividing it by 1 − .391.
h	.011	Ratio of fees and royalties to total affiliate investment. According to the U.S. Office of Business Economics (1975), Table 10, p. 49, manufacturing affiliates paid fees and royalties amounting to $1.855 billion in 1974. There is no reliable estimate of the total assets of *U.S.* manufacturing affiliates to use in

TABLE A2—*continued*

Parameter	Value	Definition and Source
		deflating this figure. I consulted the U.S. Senate Committee on Finance, Table 12, p. 432 and obtained the estimate for $78 billion in total assets in 1970. I assumed that total assets grew at an annual growth rate of 16.7 percent per annum (the rate of growth of affiliate sales) between 1970 and 1974, which gives an estimated total assets for 1974 of $170 billion. Dividing the $1.85 billion in royalties and fees by the $170 billion total assets yields the value for h of .011.
w_D, w_H, w_B	.15	Withholding tax rates applied to dividends, fees, and interest payments, respectively. I have assumed that a 15 percent rate is applied to each type of payment.
F	2.71	Net capital outflow from parent to affiliate. See U.S. Office of Business Economics (1975), Table 3, p. 47.
$R^* - B^*$	12.73	Parent's domestic income before taxes. According to Bergsten, Horst, and Moran, Table 6-1, 24 multinational manufacturers reporting sufficient statistics earned 54.4 percent of their book income before taxes from domestic operations. The $12.73 billion equals 54.4 percent of $12.73 billion plus $10.67 billion (the latter being my estimate of the affiliate's pre-tax earnings above).
r'	-.0025	The slope of the marginal revenue from new investment by the foreign affiliate. This slope was chosen because it implied that an elasticity of 2 for the marginal-revenue-from-new-affiliate-investment schedule.
$r^{*\prime}$	-.0012	The slope of the marginal revenue from new investment by the *U.S.* parent. Lacking any more reliable estimate, I assumed that the elasticity of investment demand at home, 2, was the same as that abroad.
b'	.0049	The slope of the marginal cost of outside capital schedule for the foreign affiliate Once again, lacking any better estimate I assumed a slope which would make the elasticity of the schedule equal to 2.
$b^{*\prime}$	.0024	Slope of the marginal cost of outside capital schedule for the *U.S.* parent. I chose a slope such that the elasticity would once again equal 2.
I_0	170	Foreign affiliate's total assets. See note for h above
F_0	34.	Total capital transferred from parent to affiliate. Robbins and Stobaugh, Table 4-1, Table 2, indicate that 20 percent of manufacturing affiliates' sources of funds between 1966 and 1969 came from their parents. My $34 billion estimate for F_0 is 20 percent of the $170 billion estimate for I_0.
f, f_0	.64	Ratio of intrafirm debt to total capital transfer for new and existing investment, respectively. Estimate obtained directly from U.S. Office of Business Economics (1975), Table 3, p. 47 for incorporated manufacturing affiliates.
i_p, i_{p_0}	.031	Intrafirm interest rate. James Nunns and Gary Hufbauer, Table 2, and U.S. Office of Foreign Direct Investment allows us to calculate the ratio of intrafirm interest payments to intrafirm debt This low average interest rate reflects the use of interest-free trade credits.
I	18.3	Subsidiary's new investment in 1974. The U.S. Office of Business Economics (1976) Table 1, p. 21 indicates that manufacturing affiliates capital expenditures were $11.7 billion in 1974. This figure includes only property, plant, and equipment expenditures, so I have increased this figure by 56 percent to include short-term capital formation. This 56 percent increase is based on the ratio of the estimated increase in total assets between 1970 and 1974 (see note for h above) and the total value of property, plant, and equipment expenditures over that same interval.
I^*	36.4	According to the U.S. Senate Committee on Finance, Table 12, p. 432, the increase in the total assets of the *U.S.* parents between 1966 and 1970 was 1.99 times as large as that of their foreign affiliates. My value of I^* is 1.99 times 18.3, my estimate of the value of I.

Inflationary Finance and the Dynamics of Inflation: Indonesia, 1951–72

By Bijan B. Aghevli and Mohsin S. Khan*

In studies on the role of monetary factors in hyperinflation the endogeneity of the money supply has emerged as a particularly interesting question. Phillip Cagan, in his seminal study on the dynamics of hyperinflation, treated the supply of money as an exogenous variable, whose rapid expansion caused and perpetuated the increase in prices. Recently, however, a number of studies have argued that the expansion in the money supply is itself a result of ongoing inflation, and while it is still a major factor in affecting the rate of inflation, there is evidence of two-way causation. The expansion in the nominal stock of money increases the demand for goods and services and thus prices, but the inflation results in increased government deficits which the authorities finance by further money creation. For example, in their study of seven hyperinflation cases, Thomas Sargent and Neil Wallace explicitly recognized the importance of the feedback from inflation to the expansion in the money supply by concluding: "Such feedback appears to have been present in several of the hyper-inflations that we have studied. This might be explained by the government's resorting to money creation in order to finance its expenditure," (p. 349). Similar conclusions stressing the fiscal actions of the government as contributing to inflation have been reached by Robert Barro, Rodney Jacobs (1975b), and Jacob Frenkel (1976b).[1]

The policy of financing government ex- penditures by the creation of money has been pursued by many countries. This policy has particular attraction for those govern- ments which are unable to enact adequate tax programs or administer them effectively to gain the required revenue. This form of deficit financing causes inflationary pres- sures by increasing the supply of money Moreover, as mentioned before, as the in- flation rate rises, government expenditu rises faster than revenue, forcing the au thorities to increase their issuance of mon even further. The basic reason for this sel perpetuating effect between governme deficits and prices is that nominal revenu of the government are generally fixed in th short run, and thus their real value falls i the face of rapid inflation.[2] At the same tin the government's spending commitments a mostly in real terms, which implies tha nominal expenditure rises concomitantl with price increases. Therefore even if in th long run government revenues matc government expenditures, the lag structu in the government's budgetary mechanis creates a deficit in real terms as well as i nominal terms in the presence of inflation

There have been relatively few attemp to incorporate this type of phenomeno into a model of inflation.[3] The problem wa first analyzed by Michael Lovell and the later by Julio Olivera and Dean Duttor all of whom stressed the role of the feec back between inflation and increases in th money supply. The purpose of our paper i first to develop a dynamic model of defic financing and the inflationary mechanism i a continuous time framework. The theoret

*The research on this paper was partially funded with a grant from the United Kingdom Social Science Research Council while we were at the International Monetary Research Programme, London School of Economics, on leave from the International Monetary Fund. We wish to thank G. Borts, R. Findlay, J. A. Frenkel, H. G. Johnson, P. R. Narvekar, I. Otani, and C. R. Wymer for many valuable comments and suggestions. The suggestions of an anonymous referee led to a substantial improvement in the paper. The views expressed and any remaining errors are our own.

[1]This aspect has also been acknowledged recently by Cagan. See Cagan and George Kincaid.

[2]This would cover revenues from sales taxes, prop erty taxes, and income taxes that are paid at their a sessed nominal values.

[3]The studies of Sargent and Wallace and Frenk (1976b) contain empirical evidence on the relationsh between inflation and changes in the money supply

TABLE 1—INDONESIA: MONETARY AND BUDGETARY STATISTICS, 1952-72
(shown in percent)

	Rate of Growth of Money Supply	Rate of Inflation	Ratio of Government Expenditure to Income	Ratio of Government Revenue to Income
1952	13.6	22.2	17.4	14.3
1953	19.8	0.0	18.4	15.9
1954	26.0	10.4	17.0	12.9
1955	24.6	25.6	13.3	11.6
1956	9.3	19.8	14.7	13.6
1957	21.6	21.1	15.6	12.4
1958	39.3	30.2	13.5	9.1
1959	30.6	14.4	15.6	8.2
1960	24.4	15.1	14.9	12.8
1961	33.3	39.3	18.7	13.2
1962	52.0	80.1	9.1	5.5
1963	68.2	88.9	10.3	5.0
1964	80.4	57.7	9.5	4.1
1965	113.5	153.1	10.6	3.9
1966	174.7	234.7	9.3	4.1
1967	149.9	137.7	10.3	7.1
1968	81.9	68.5	8.9	7.6
1969	63.5	35.4	10.1	8.7
1970	39.2	9.0	13.0	9.6
1971	28.0	5.5	13.9	10.3
1972	32.0	12.7	14.7	10.1

Sources: Bank of Indonesia (1972, 1973); World Bank; Sundrum (1972).

model is a disequilibrium system for-ılated in terms of a set of differential ıations, where real money balances, ɾernment expenditure and revenue adjust h a lag to their desired levels. Our for-ılation allows us to estimate the lag ucture of the dynamic system directly tead of imposing the lags arbitrarily, as previous studies. For example, both ıvera and Dutton developed discrete time ıdels where all the lags are fixed a priori. ıce the key element in the analysis is the ıerent lag structure, the imposition of ar-rary lags is less than satisfactory.[4] Using nual data for the period 1951-72, the ıdel is then estimated for Indonesia in ntinuous time, applying the methods developed by J. D. Sargan (1974) and C. R. Wymer (1972, 1976). One of the advantages of this estimation method is that the estimator is independent of the observation period so that the model can be specified and analyzed independently of the sample being used for estimation.

In Section I we describe the individual equations that make up the complete model. The results from estimating this model, along with a simulation of the complete model, are discussed in Section II. The stability properties of the estimated model are examined in Section III, and in Section IV we consider some relevant policy issues of deficit financing and the welfare costs of inflation. The implications of our analysis are contained in the concluding section.

[4]Insofar as Olivera was more interested in bringing ıt the basic features of the self-perpetuating nature of ılation and less with empirical verification of the eory, the limitations of his formulation are not seri-s. Dutton, however, is mainly interested in testing s model, and therefore imposing the lags a priori is ıt justified.

I. A Dynamic Model of Inflation

Before formulating the model it would be useful to review in broad outline the monetary and price developments of the *In-*

donesian economy over the period 1951–72 (see Table 1) and some of the factors underlying these developments.[5] The experience of the Indonesian economy during the last two decades provides an interesting modern case study for analyzing the dynamics of inflation. The relative price stability achieved by the economic reforms of the mid-1950's was soon undermined by the rapid monetary expansion resulting from the financing of military operations during the Sumatran rebellion of 1957–58. Despite the substantial price increases during 1958–61 and the resulting cost overruns, the authorities not only maintained but substantially increased their total expenditure in real terms. The year 1961 saw the start of the ambitious eight-year development plan and preparations for the West Irian campaign. The pressure on prices was also greatly exacerbated by the severe drought of that year. Government expenditures were more than doubled during 1960–62 in nominal terms, but declined substantially in real terms. In the subsequent years of hyperinflation, the government again increased expenditures in real terms, even as revenues in real terms declined steeply. As we describe later, the decline in revenues was a direct consequence of high inflation, given the structure of taxation and the method of collection of taxes in Indonesia during the observation period.

The main elements in the dynamics of the accelerating phase of Indonesian inflation emerge clearly from the above description. Price increases were initiated, and began to be aggravated, by rising government expenditures in the late 1950's and by the effect of the drought in 1961 on food supplies and prices. The rising prices themselves led to sharp increases in nominal government expenditures. Since government revenues lagged substantially behind price developments, the authorities were forced to finance their deficits by the creation of money. In turn, the rise in the money supply reinforced the inflationary process and prices accelerated even more. This self perpetuating process caused inflation to spiral higher and higher until it reached the stage of near hyperinflation in the mid 1960's.[6]

We will now develop a model of the inflationary process for the Indonesian case. The approach taken is that inflation in Indonesia is basically a monetary phenomenon,[7] and that the monetary expansion itself is linked to the rate of inflation through the government budget, and is therefore treated as endogenous.[8]

A. *Demand for Money*

The demand for real money balances is specified as a function of real income and the expected rate of inflation, with the latter variable being used as a proxy for the opportunity cost of holding money. In Indonesia, as well as in most other developing countries, capital markets are not well developed, and thus the alternative to holding money balances is for the most part to hold goods. Since the real value of money balances is decreased by rises in the price level, the expected rate of inflation can be treated as the opportunity cost of holding money. There is, of course, some substitutability between money and bonds, and a rate of interest should also be introduced to measure the opportunity cost of holding money relative to bonds. Rates of interest in Indonesia, however, are controlled by the authorities and thus not accurate in reflecting market conditions. On pragmatic grounds we have excluded the rate of interest from the relationship.

[5]Excellent background material can be found in the papers by H. W. Arndt and R. M. Sundrum

[6]Strictly speaking, the Indonesian inflation never reached the hyperinflation stage as defined by Cagan (i.e., 600 percent or more per annum). Cagan's definition is rather arbitrary, however, and one could argue that the Indonesian inflation of the mid-1960's shared most of the common characteristics of other hyperinflation cases.

[7]This appears to be consistent with other studies of inflation in developing countries. See, for example, Arnold Harberger, Adolpho Diz, Colin Campbell, and more recently, Robert Vogel.

[8]This is in contrast to previous studies where monetary expansion is considered to be the main exogenous factor leading to a rise in prices.

The desired stock of real money balances is formulated as follows:[9]

(1) $(M/P)^d = aYe^{-b\pi} \quad b > 0$

where M/P = stock of real money balances
Y = level of real income
P = price level
π = expected rate of inflation

The income elasticity of the demand for money is constrained to equal unity.[10] The parameter a is the inverse of velocity consistent with zero rate of inflation and b is the inflation coefficient.

In this framework, it is assumed that the authorities control the nominal stock of money while the public determines the real stock with prices adjusting to clear any disequilibrium in the money market. It is further assumed that the public does not adjust its balances instantaneously. The rate by which the actual balances are adjusted to desired levels is a function of excess demand for money and this relationship is written according to the following log-linear function, where D is a differential operator ($D = d/dt$):

$$(2) \quad D\,log\,(M/P) = log\left[\frac{(M/P)^d}{(M/P)}\right]^{\delta}$$
$$= \delta[log\,a + log\,Y - b\pi - log\,(M/P)] \quad \delta > 0$$

In this formulation, the partial adjustment coefficient δ measures the speed of adjustment. That is, the inverse of δ is the mean lag of the adjustment process (it takes a period of $1/\delta$ for 66.6 percent of any excess demand for money to be eliminated through the change in real balances).

Inflationary expectations are formed according to the following adaptive expectations formulation:[11]

$$(3) \quad D\pi = \phi[D\,log\,P - \pi]$$

which implies $\pi = \dfrac{\phi}{D + \phi} D\,log\,P$

Substituting for π from (3) into (2) and rearranging results in the following second-order differential equation:

$$(4) \quad D^2\,log\,(M/P) + \phi\,D\,log\,(M/P) = \delta[a\phi + D\,log\,Y + \phi\,log\,Y - b\phi\,D\,log\,P - b\,D\,log\,(M/P) - \phi\,log\,(M/P)]$$

In principle the above formulation allows one to differentiate between the speed of adjustment of real balances to their desired level δ, and the speed of adjustment of inflation expectations ϕ, by estimating both parameters separately. In order to estimate second-order systems of differential equations reliably, however, one needs fairly accurate monthly or at least quarterly data. In the absence of such data, we are forced to approximate the above second-order differential equation by a first-order one where both adjustment parameters of real balances and inflation expectations are incorporated into one parameter. As Cagan has pointed out, one can assume that either real balances or inflation expectations adjust instantaneously (i.e., either δ or ϕ approaches infinity) and derive the identical reduced form for the rate of inflation given below:[12]

[9] In the absence of clear evidence on whether the appropriate functional form is double-log or semilog, and because most studies have used the latter specification, we have maintained the semilog form. For a discussion of this issue see Frenkel (1976b).

[10] In order to test the validity of this assumption, the income elasticity of the demand for money was first estimated using a single equation approach. This resulted in a point estimate of 1.02 which was not significantly different from unity. We then constrained the value of income elasticity to one in order to increase the efficiency of the other estimates of the model.

[11] In a more general framework, one should probably allow for a regressive element in expectation formations along the lines suggested by Frenkel (1975) and Michael Mussa. In a simultaneous model, however, this would complicate the estimation process substantially, and we have thus neglected the regressive element in the formation of inflation expectations.

[12] Assuming that there is no lag in formation of inflation expectations (i.e., $\pi = D\,log\,P$) and manipulating equation (2) results in equation (5) where λ is equal to δ. Alternatively, assuming there is no lag in adjustment of real balances (i.e., $(M/P)^d = (M/P)$) and substituting from (3) into (1) results in the identical reduced form as equation (5) where $\lambda = \phi$ with only one additional term in the bracket which is $D\,log\,Y$. The rate of growth of real income in a situation of high inflation is quite negligible relative to the other terms in equation (5) and it can be ignored.

$$(5)\quad D\,log\,P = \frac{1}{1-\lambda b}\,[D\,log\,M + \lambda\,log\,M - \lambda\,log\,P - \lambda\,log\,Y - \lambda\,log\,a]$$

In this formulation, the parameter λ is a composite coefficient which can be interpreted as a mixture of both lags involved in adjustment of real balances as well as inflation expectations. Some recent empirical work by Teh-wei Hu, who has estimated a similar relationship to equation (4) for the Chinese hyperinflation, indicates that, at least for that case, real balances adjusted to the desired level very rapidly while the lag involved in the adjustment of inflation expectations was considerably longer. This finding seems to be in line with the intuitive expectations regarding the relative importance of the two lags involved. Thus, we suspect that the coefficient λ in equation (5) will measure mostly the adjustment coefficient ϕ of the formation of inflation expectations.

One point to note about equation (5) is that since it specifies inflation as a dependent variable it avoids any of the problems with Cagan-type money demand functions that Jacobs (1975a) has recently mentioned. Equation (4), which is essentially the estimating form used by Dutton, may yield biased parameters due to spurious correlation between prices and nominal money balances.[13]

B. *Government Budget*

In Indonesia, as in most developing countries, the level of private investment is regarded by the government authorities as generally too small to provide the economy with adequate capital formation. The government is itself thus forced to finance many large development projects in order to meet certain growth targets. This would imply that the authorities are committed to meet certain expenditures regardless of cost overruns. Assume the authorities desire to fix the real value of the expenditures they undertake as a constant fraction of real income. Due to many institutional bottlenecks, the level of real government expenditure cannot be adjusted to the desired level instantaneously. Assuming that the real expenditure is adjusted to the desired level with a lag, the following relationship can be written where the rate of change of real expenditure G is a log-linear function of the discrepancy between the actual and desired levels, where G^d denotes the desired level of expenditure.

$$(6)\quad D\,log\,G = log\left[\frac{G^d}{G}\right]^{\gamma} = \gamma[log\,g + log\,Y - log\,G]$$

where $\quad G^d = g\,Y$

Similarly, the desired level of government revenue can be assumed to be a constant fraction of income. Due to the structure of taxation and the method of collecting taxes in Indonesia, however, nominal revenue tended to adjust slowly to price developments. The Indonesian tax system depended heavily on indirect taxes and, in particular on foreign trade taxes. The buoyancy in tax revenues that obtains during periods of rising nominal income in economies with progressive direct tax systems was therefore not in evidence in Indonesia. The indirect taxation of domestic transactions was in large part either specific or ad valorem with infrequently adjusted base values. It was therefore not adequately responsive to changes in the nominal values of transactions. Also, the normal lag in collecting taxes meant, in the period of inflation, a substantial lowering of the real value of tax collections. Further, the efficiency of the tax collection machinery was seriously affected by the inflation. Proceeds from the taxation of exports fell, not only because of the decline in the world market prices of rubber and other principal items that occurred during the first half of the 1960's, but also because of a reduction in exportable surpluses

[13] Jacobs (1975a) actually proposes relating real money balances to current and past values of the nominal money stock rather than current and past value of prices as is done in standard Cagan-type models. Since in such models it is the rate of inflation that is the endogenous variable, equation (5) would seem to be more appropriate. For a discussion of this issue see Khan and also Cagan and Kincaid.

and large-scale underinvoicing of exports. With inadequate adjustment of the exchange rate, the real value of import duty collections also declined. Nevertheless, as the actual taxes fell short of the desired levels, the authorities put more pressure on the tax collection agencies to increase the tax revenues through more "tax effort." The rate of change of nominal taxes can thus be specified analogous to the expenditure equation where T^d denotes the desired level of nominal taxes.

$$(7) \quad D \log T = \log \left[\frac{T^d}{T}\right]^\tau = \tau[\log t + \log Y + \log P - \log T]$$

where $\quad T^d = tYP$

There are two a priori observations which can be made in regard to the coefficients of equations (6) and (7). First, one would expect that the authorities would be able to adjust the level of real expenditure to the desired level faster than nominal taxes. Second, the value of the authorities' desired propensity to tax t, should be close to the desired propensity to spend, g. Therefore, the parameters would have the following properties: $0 < \tau < \gamma < \infty$; and $t \simeq g$.

C. *Money Supply*

Under a fractional reserve system, the change in the money supply can be specified as a multiple of the change in the stock of reserve money:

$$(8) \quad D \log M = DM/M = \eta DRM/M$$

where η is the fractional reserve coefficient (the money multiplier) and RM is the stock of reserve money. The money multiplier was relatively stable over the period and we have chosen to treat it as a fixed parameter.[14]

Changes in reserve money come from three sources: the government domestic deficit; changes in central bank credit to the nongovernment sector; and the private sector's balance of payments. The private sector's balance of payments was clearly adversely affected by the rapid inflation and the inappropriateness of the exchange rate. Thus, the deficit in the balance of payments tended to offset partially the effect of the fiscal deficit on the reserve money, particularly in the early period. As the rate of inflation increased, however, the role of balance of payments in determination of reserve money was minimized as the authorities relied on frequent devaluations and severe trade controls in order to counteract the effects of domestic inflation. In a more complete model the balance of payments can also be studied, but for this study it is assumed to be exogenous. This assumption should not alter the results too much as most of the changes in reserve money have originated in government deficit financing.

The changes in reserve money can be written according to the following identity:

$$(9) \quad DRM = GP - T + H$$

H is a residual item that includes the private sector's balance of payments, foreign expenditures made by the government, and changes in claims of the central bank on the nongovernment sector. In this model H is treated as an exogenous variable. Substituting equation (9) into (8) results in the following equation for the money supply:

$$(10) \quad D \log M = \eta \left[\frac{GP - T + H}{M}\right]$$

II. Estimation Results

The model consists of four first-order differential equations determining the rates of change of prices, real government expenditure, nominal taxes, and money supply given by equations (5), (6), (7), and (10), respectively. The complete model is reproduced below:

$$(11a) \quad D \log P = \frac{1}{1 - \lambda b} [D \log M + \lambda \log M - \lambda \log P - \lambda \log Y - \lambda \log a]$$

[14] Since banking development in Indonesia was in relatively early stages, the money supply was mostly comprised of currency, and the banks did not engage in large amounts of multiple creation of money under the fractional reserve system. Thus, the value of money multiplier remained around 1.2 over this period with relatively smaller fluctuations. For an analysis of components of the money multiplier, see Aghevli (1977a).

$$\text{(11b)} \quad D \log G = \gamma[\log g + \log Y - \log G]$$

$$\text{(11c)} \quad D \log T = \tau[\log t + \log Y + \log P - \log T]$$

$$\text{(11d)} \quad D \log M = \eta\left[\frac{GP - T + H}{M}\right]$$

The first three equations are linear in logarithms whereas the last equation, determining the rate of change of money supply, is non-linear, thus making the complete model non-linear in variables. Since for estimation by any systems method it is much easier to work with linear models, the last equation was linearized by taking its Taylor series expansion around its steady-state values.[15]

The complete model was estimated in continuous time using the technique developed by Sargan, and Wymer (1972). This technique is particularly suitable for estimating dynamic systems as it allows one to specify the model as a set of differential equations. Conventional econometric models are usually specified in terms of discrete time, which imposes the entirely arbitrary assumption that the decision period of all economic agents is the same and that the frequency of decisions happens to correspond to the aggregate data available. The estimation method used here provides, first, a discrete time approximation to the continuous model; and second, since the approximation introduces a moving average process into the error structure, efficient and consistent estimates of the parameters in the approximated model.[16]

Since the model is linearized we were able to estimate it using the Full Information Maximum Likelihood (*FIML*) estimator—*FIML* utilizes all a priori restrictions on the complete system in order to estimate the structural coefficients simultaneously by maximizing the likelihood function of the model. On the basis of asymptotic theory it appears to be the best estimator.

The estimates of the structural model are given in Table 2, where the standard errors of the parameters are given in parentheses below the respective coefficients. (For data sources, see Table 1.)

All the parameters in the model have the expected signs and are significantly different from zero at the 5 percent level. The root

TABLE 2—MODEL ESTIMATES

Rate of inflation

$$D \log P = \underset{(0.20)}{1.91}[D \log M + \underset{(0.09)}{0.87}(\underset{(0.07)}{-2.34} + \log M - \log P - \log Y)]$$

$RMSE = 0.28$

Government expenditure

$$D \log G = \underset{(3.60)}{8.03}[\underset{(0.05)}{2.09} + \log Y - \log G]$$

$RMSE = 0.26$

Government revenue

$$D \log T = \underset{(0.52)}{2.64}[\underset{(0.08)}{2.29} + \log Y + \log P - \log T]$$

$RMSE = 0.20$

Money supply

$$D \log M = \underset{(0.06)}{1.18}\left[\frac{GP - T + H}{M}\right]$$

$RMSE = 0.13$

[15]This linearization procedure about the steady state yields the equation:

$$D \log M = -m_0 - m_1(\log G + \log P) - m_2 \log T - \log M - m_4 \log H$$

where

$$m_0 = -\eta\left[\frac{P^*G^* - T^* + H^*}{M^*}\right] + \eta\frac{P^*G^*}{M^*} \cdot \log\left[\frac{P^*G^*}{M^*}\right] - \eta\frac{T^*}{M^*}\log\left[\frac{T^*}{M^*}\right] + \eta\frac{H^*}{M^*}\log\left[\frac{H^*}{M^*}\right]$$

$$m_1 = -\eta\frac{P^*G^*}{M^*}; \quad m_2 = \eta\frac{T^*}{M^*}$$

$$m_3 = \eta\left[\frac{P^*G^* - T^* + H^*}{M^*}\right]; \quad m_4 = -\eta\frac{H^*}{M^*}$$

where P^* refers to the steady-state value of P and so on.

[16]In certain special cases when the model comprises only stock or instantaneous variables, such as the Cagan model, the discrete approximation may not cause any problems with the errors. However, in any general mixed stock-flow model the approximation procedure will result in the errors containing a moving average. For a full discussion of the econometric issues involved the reader is referred to Sargan, and Wymer (1972, 1976).

TABLE 3—STRUCTURAL PARAMETER ESTIMATES

$\lambda = 0.87$	$\eta = 1.18$	$a = 0.097$
(0.09)	(0.06)	(0.007)
$\gamma = 8.04$	$g = 0.123$	$b = 0.549$
(3.65)	(0.006)	(0.079)
$\tau = 2.64$	$t = 0.102$	
(0.53)	(0.008)	

mean-squared error of each equation is denoted by *RMSE* and represents the percentage error in the level of the endogenous variables.[17] Based on the above estimated coefficients, the values of the parameters of the system are given in Table 3, where the standard errors are given in parentheses underneath the coefficients. In the first equation the constant term implies that the desired velocity of money consistent with zero inflation is about 10, and the inflation elasticity is not inconsistent with the values obtained by Cagan (1956).[18] The partial adjustment coefficients for inflation, real government expenditure, and taxes imply mean time lags of 14, 1.5, and 4.5 months, respectively. The mean time lag for taxes is about three times as large as the government expenditure lag, thus indicating that in a period of rising prices the revenue from taxes will continually fall short of government expenditures and result in increasing deficits. It should be noted that the coefficients g and t are not significantly different from each other, thus implying that the marginal income propensities of desired government expenditure and taxes are almost the same. Therefore, even though the authorities set the targets of expenditure and revenue from taxes to match, an increase in inflation will still produce a deficit due to the presence of lags in the system. The estimated coefficient for the money multiplier η implies that government deficits affect the money supply on an almost one-to-one basis.

In order to test the goodness-of-fit of the estimated model shown in Table 2, we performed a within-sample dynamic simulation for the four endogenous variables. A comparison of the actual and simulated values gives an indication of whether the model is able to capture the historical behavior of the endogenous variables.[19] The simulated and actual values (*logs*) of prices, real government expenditures, government revenue from taxes, and the money supply, are shown in Figure 1. The charts show that the model is fairly accurate in capturing the movements of the four variables. The simulated values, especially towards the later part of the period, lie very close to the actual values. Table 4 represents the correlation coefficients and root mean-squared errors between the actual and simulated values of the four endogenous variables. Both sets of statistics provide support for the goodness-of-fits observed in Figure 1. There is no indication of any explosive behavior on the part of the simulated values since the deviations between the actual and simulated paths of the variables grow smaller over time.[20]

TABLE 4—COMPARISON OF ACTUAL AND SIMULATED VALUES

Variable	Correlation Coefficient	Root Mean-Squared Error
Prices	0.997	0.068
Government expenditure	0.998	0.062
Government revenue	0.998	0.047
Money supply	0.999	0.031

III. Dynamic Stability of the Model

The question of stability and steady-state equilibrium is of particular interest for a

[17]It should be noted that the root mean-squared ors refer to the *rates of changes* of the dependent iables

[18]When Cagan values are converted to an annual sis.

[19]A dynamic simulation can also be used to test the stability of the estimated model by examining the deviations between the actual and simulated paths of the endogenous variables. In the following section this question of stability will be considered in more detail.

[20]It should be noted that some further work on the Indonesian monetary sector seems to indicate that there may have been some structural changes in the parameters of the system following the change of government in 1966. Due to the lack of adequate observation, however, we were not able to test for this change rigorously.

FIGURE 1. INDONESIA
DYNAMIC SIMULATION OF THE MODEL, 1954–72

model of the type described in this paper. As stated earlier, as prices rise, nominal expenditure rises faster than nominal revenue. Financing this "excess expenditure" will cause an increase in the money supply, raising prices further. It is important to analyze whether such a self-perpetuating process of inflation converges to a steady-state equilibrium value or whether the system is explosive.

One way to analyze stability is to give the dynamic system an exogenous shock and trace its effects on the endogenous variables far into the future until one is satisfied that the system is not explosive. This type of simulation exercise has been carried out by Dutton in order to establish the stability of his model. Like a dynamic simulation this procedure, however, is merely suggestive since it will not prove stability conclusively. In our model the issue of stability can easily be studied by analyzing the eigensystem of the dynamic model. Disregarding the movement of the exogenous variables, a necessary and sufficient condition for stability is that the real parts of the eigenvalues of the

TABLE 5—ESTIMATES OF EIGENVALUES OF THE SYSTEM

Eigenvalues			
Real Part	Imaginary Part	Damping Period (Months)	Period of Cycle (Months)
−0.82 (0.09)		14.5	
−8.04 (3.60)		1.5	
−2.00 (1.03)	±2.12 (0.37)	6.0	35.5

system have negative values. An explicit derivation of the eigenvalue of the system is too cumbersome but the eigenvalues can be estimated numerically from the estimated coefficients. The calculated values of the eigenvectors are given in Table 5, where the standard errors are given underneath the coefficients. As it can be seen the system has two real roots and two complex roots which are complex conjugates of each other, and since all the roots have negative real parts, the system is stable.

The next issue is to derive the steady-state value of the inflation rate that would prevail if no exogenous shocks were given to the system. Let ρ and μ denote the steady-state rates of growth of the money supply and real output, respectively. It can be shown that the steady-state level of real money balances, real government expenditure, and nominal taxes are given by the following relationships.[21]

$$(12)\quad (M/P) = aY\exp\left[-\frac{\mu}{\lambda} - b(\rho-\mu)\right]$$

$$(13)\quad G = gY\exp\left[-\frac{\mu}{\gamma}\right]$$

$$(14)\quad T = tYP\exp\left[-\frac{\rho}{\tau}\right]$$

It would be useful to present some of the steady-state implications of the model. In the steady state, rate of monetary expansion ρ would be equal to the real rate of growth μ, plus the rate of inflation π (i.e., $\rho = \mu + \pi$). Utilizing this relationship we can derive the following expression which specifies the real deficit as a function of inflation and speeds of adjustment of government expenditure and revenue:

$$\log\left(\frac{GP}{T}\right) = \log(g/t) + \mu\left(\frac{1}{\tau} - \frac{1}{\gamma}\right) + \frac{\pi}{\tau}\quad .$$

The above relationship indicates that for given speeds of adjustment, τ and γ, higher rates of inflation induce higher real deficits. This property of the model should come as no surprise as the model was constructed in order to precisely capture this element of the self-perpetuating mechanism. Moreover, a lower speed of adjustment of taxes τ, and a higher speed of adjustment of government expenditure also lead to higher real deficits as expected. It should be noted that the steady-state inflation rate itself is endogenous in our model and can be derived as a function of the various speeds of adjustment of the dynamic model. Substituting equations (12), (13), and (14) into (11d) results in

$$(15)\quad \rho = (g/a)\exp\left[\frac{\mu}{\lambda} - \frac{\mu}{\gamma} + b(\rho-\mu)\right] - (t/a)\exp\left[\frac{\mu}{\lambda} - \frac{\rho}{\tau} + b(\rho-\mu)\right]$$

The steady-state rate of growth of real income μ, can be approximated by the average annual rate of growth over the period. The value of μ is about 0.016 and it remained relatively fixed over the period under study. Substituting the numerical values

[21] In order to derive the steady-state level of nominal balances, real government expenditure, taxes, and inflation, consider the following solutions to the dynamic system made up of (11a), (11b), (11c), and (11d), where ρ_i denotes the steady-state rate of growth in each equation, μ is the exogenously given rate of growth of output, and t is a time trend:

$$M = M_0\exp(\rho_1 t)$$
$$G = G_0\exp(\rho_2 t)$$
$$T = T_0\exp(\rho_3 t)$$
$$P = P_0\exp(\rho_4 t)$$

Substituting the above equations into the system and setting the exogenous variable H equal to zero, we get equations (12), (13), and (14) as a steady-state solution of the system. It should be noted, however, that this solution is not necessarily unique.

for the parameters we can solve equation (15) numerically to get the steady-state rate of monetary expansion ρ and inflation (i.e., $\pi = \rho - \mu$). The value of ρ is about 165 percent, which implies a steady-state rate of inflation of about 163 percent. Such a high steady-state rate of inflation points to the inherent tendency of the system to generate high inflation rates. Since the government's desired propensity to spend in our results was very close to its desired propensity to tax, the system's tendency to generate high rates of inflation is mostly a consequence of the longer lag in the adjustment of taxes.

IV. Welfare Costs and Revenue from Inflation

In this section we will consider some of the implications of the deficit financing policies pursued by the authorities. As pointed out by Cagan and others, deficit financing can be viewed as a tax on holders of real balances whose assets are continuously devalued by the inflation. The government revenue from inflation tax can then be written as follows where rate of monetary expansion ρ, and the value of real balances (M/P) can be viewed as the "tax rate" and the "tax base," correspondingly.

$$\frac{\dot{M}}{P} = \rho \left(\frac{M}{P}\right) \tag{16}$$

In this formulation higher rates of monetary expansion lead to a reduction of tax base due to the induced inflation which reduces the demand for real balances. The value of inflation tax will thus be maximized at the point where the marginal rise due to the increase in the tax rate is equal to the marginal fall due to the decrease in the tax base. Thus, the revenue maximizing rate of monetary expansion is given by the expression $\rho = 1/b$.[22] Substituting the estimated value of b, which is 0.55, results in a value of 180 percent for the revenue-maximizing rate of monetary expansion.[23]

It is interesting to note that this value is very close to the steady-state rate of monetary expansion of the model. Thus, it seems that the authorities' budgetary policies were consistent with maximizing the government revenue from inflation tax, which led to very high rates of inflation. High rates of inflation, however, reduced the real value of other taxes due to collection lags involved, as can be seen from equation (14). Moreover, high inflation rates impose additional costs on the public which can be viewed as a cost of collection for the inflation tax.

There are two types of cost associated with inflation. First, inflation affects the distribution of resources. This effect will be particularly pronounced in the short run since the unanticipated inflation rate affects the various sectors of the economy differently. The distributional impact of inflation is quite complex and depends on many institutional factors which cannot be readily quantified. For the purposes of this study, the redistributive-disruptive effects of inflation are neglected by assuming that inflation is fully anticipated and discounted for by the government and the public. The second cost of inflation is the welfare cost of the reduction in real balances due to inflation. Following the literature on this topic,[24] inflation can be viewed as a tax on holders of real balances. This cost can be computed in the steady state by measuring the area under the demand for real balances curve as suggested by Bailey. To use Bailey's analysis, the inflation can be viewed as "the subjective marginal rate of substitution of

[22] It is clear from expression (17) that the revenue from inflationary finance is maximized by setting ρ equal to $1/b$. See Robert Mundell, and Alvin Marty (1967, 1973).

[23] Our value of the revenue maximizing rate of monetary expansion is much higher than the values obtained by Milton Friedman and Frenkel (1976a). The major difference seems to be in Friedman's assumed values for the inflation elasticity b (i.e., 2, 10, and 20) which were also subsequently used by Frenkel, and our estimated value of 0.55. It is interesting to note that our estimated value of b (i.e., 0.55) is quite close to the values provided by Cagan for a set of hyperinflation countries, Campbell for Korea and Brazil, Diz for Argentina, and Vogel for a set of Latin American countries.

[24] For an analysis of this concept see Martin Bailey and Marty (1967, 1973). It should be mentioned that for simplicity, we have only considered the welfare cost of inflation in the steady state. The considerations of dynamic paths, however, could alter the results as the welfare cost of inflation is quite sensitive to the formation of inflation expectations. See Leonardo Auernheimer, Charles Cathcart, and Frenkel (1975a).

TABLE 6—REVENUE MAXIMIZING AND PRICE STABILIZING RATES OF MONETARY EXPANSION

Rate of Monetary Expansion in Percentage ρ	Tax Revenue as Percent of Income T/YP	Revenue from Deficit Financing as Percent of Income M/YP	Welfare Cost of Inflation as Percent of Income W
$\rho = \mu = 1.6$	10.1	0	0
$\rho = 1/b = 182$	5.0	6.0	14

eal goods for cash balances for everyone olding the latter" (p. 94). The welfare cost f inflation as a fraction of income can then e determined by the following integral where m_0 and m_ρ correspond to the levels of eal balances consistent with rate of monetary expansion of zero and ρ, respectively.[25]

$$(17)\quad W = \frac{1}{Y}\int_{m_0}^{m_\rho} \pi d(M/P) = \frac{a}{b} - \left(\rho - \mu + \frac{1}{b}\right) a \exp[-b(\rho - \mu)]$$

n order to have some idea of the relative mportance of this welfare cost, we have computed the revenue from normal taxes in addition to the revenue from deficit financing and the welfare cost of the induced inflation, as ratios of income from expressions (14), (16), and (17). The results are given for he revenue maximizing rate of monetary expansion (i.e., $\rho = 1/b$) as well as the price stabilizing rate of monetary expansion (i.e., $\rho = \mu$) in Table 6. A comparison of the two sets of computations reveals that if the authorities choose to maximize their revenue from deficit financing instead of pursuing the goal of price stability, they will increase their total revenue from 10 to 11 percent of the national income. This increase is quite marginal since much of their rise in revenue from deficit financing is counterbalanced by the fall in the value of other taxes due to he collection lags which increase with the nflation rate.[26] Moreover, this marginal rise in the revenue is achieved by imposing a substantial welfare cost on the public. In fact, the welfare cost of inflation at the revenue maximizing rate of monetary expansion is larger than the total revenue collected by the authorities. Thus, it is clear that it does not make much sense to indulge in such expansionary policies. In fact, the authorities would have been able to secure just as much real revenue had they pursued a policy of price stability by reducing the lag in collection of taxes.

[25]Substituting from equation (12) and integrating by parts results in expression (17).

[26]In the absence of alternative means to obtain revenue, however, a case can be made for moderate amounts of deficit financing when the proceeds are used to finance development expenditure; see Aghevli (1977b)

V. Conclusion

In this paper we have developed a dynamic model of inflation based on the idea that the rate of inflation tends to increase nominal expenditure faster than revenue. The resultant budget deficit increases the money supply and induces further inflationary pressures. This self-perpetuating process is formulated in a continuous time framework and a system of stochastic differential equations is estimated simultaneously. This model seems to explain the Indonesian inflation quite well.

The model was then used to draw some policy implications in regard to the steady-state rate of monetary expansion which would maximize government revenue. It was shown that while high rates of monetary expansion would increase the government revenue from issuance of money, the resultant increase in inflation would reduce the real value of taxes which adjust with a lag to price developments, leading only to marginal increases in total government revenues. Moreover, higher rates of inflation impose a substantial welfare cost which can be viewed as a collection cost of inflation tax. Incorporating all of these ele-

ments, it was argued that the authorities should aim to keep the goal of price stability by increasing the speed of adjustment in their tax collection which would break the vicious cycle of the self-perpetuating inflation.

This study also has important implications for other countries which resort to deficit financing indiscriminately. As the Indonesian case indicates, the self-perpetuating process of inflation could easily lead to hyperinflation causing serious economic, as well as political, instability.

REFERENCES

B. B. Aghevli, (1977a) "Money, Prices and the Balance of Payments," *J. Develop. Stud.*, 1977 forthcoming.

———, (1977b) "Inflationary Finance and Growth," *J. Polit. Econ.*, 1977 forthcoming.

H. W. Arndt, "Banking in Hyperinflation," *Bull Indonesian Econ. Stud.*, Oct. 1965, *1*, 45–70.

L. Auernheimer, "The Honest Government's Guide to the Revenue from the Creation of Money," *J. Polit. Econ.*, May/June 1974, *32*, 598–606.

M. J. Bailey, "The Welfare Cost of Inflationary Finance," *J. Polit. Econ*, Apr. 1956, *64*, 93–110.

R. J. Barro, "Inflationary Finance and the Welfare Cost of Inflation," *J. Polit. Econ.*, Sept./Oct. 1972, *80*, 978–1001.

P. Cagan, "The Monetary Dynamics of Hyperinflation," in Milton Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

——— **and G. Kincaid,** "Jacobs' Estimates of the Hyperinflation Model—A Comment," *Econ. Inq.*, 1977 forthcoming.

C. D. Campbell, "The Velocity of Money and the Rate of Inflation: Recent Experiences in South Korea and Brazil," in David Meiselman, ed., *Varieties of Monetary Experience*, Chicago 1970.

C. D. Cathcart, "Monetary Dynamics and the Efficiency of Inflationary Finance," *J. Money, Credit, Banking*, May 1974, *6*, 169–90.

A. C. Diz, "Money and Prices in Argentina, 1935–1962," in David Meiselman, ed., *Varieties of Monetary Experience*, Chicago 1970.

D. S. Dutton, "A Model of Self-Generating Inflation," *J. Money, Credit, Banking*, May 1971, *3*, 245–62.

J. A. Frenkel, "Inflation and the Formation of Expectations," *J. Monet. Econ.*, Oct. 1975, *1*, 403–21.

———, (1976a) "Some Dynamic Aspects of the Welfare Cost of Inflationary Finance," in Ronald I. McKinnon, ed., *Money and Finance in Economic Growth and Development: Essays in Honor of E. S. Shaw*, New York 1976.

———, (1976b) "The Forward Exchange Rate, Expectations and the Demand for Money: The German Hyperinflation," unpublished paper, Mar. 1976.

M. Friedman, "Government Revenue from Inflation," *J. Polit. Econ.*, July/Aug 1971, *79*, 846–56.

A. C. Harberger, "The Dynamics of Inflation in Chile," in Carl F. Christ, ed., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford 1963.

T. W. Hu, "Hyperinflation and the Dynamics of the Demand for Money in China, 1945–49," *J. Polit. Econ.*, Jan. 1971, *79*, 186–95.

R. L. Jacobs, (1975a) "A Difficulty w[illegible] Monetarist Models of Hyperinflatio[illegible] *Econ. Inq.*, Sept. 1975, *13*, 337–60.

———, (1975b) "Hyperinflation and [illegible] Supply of Money," unpub., 1975.

M. S. Khan, "The Monetary Dynamics [illegible] Hyperinflation: A Reply," *J. Mon[illegible] Econ.*, 1977 forthcoming.

M. C. Lovell, "A Keynesian Analysis [illegible] Forced Savings," *Int. Econ. Rev.*, Se[illegible] 1963, *4*, 247–64.

A. L. Marty, "Growth, Satiety and the T[illegible] Revenue from Money Creation," *J. Po[illegible] Econ.*, Sept./Oct. 1973, *81*, 1136–52.

———, "Growth and the Welfare Cost [illegible] Inflationary Finance," *J. Polit. Eco[illegible]* Feb. 1967, *75*, 71–76.

R. A. Mundell, "Growth, Stability and [illegible]

flationary Finance," *J. Polit. Econ.*, Apr. 1965, *73*, 97–109.

M. Mussa, "Adaptive and Regressive Expectations in a Rational Model of the Inflationary Process," *J. Monet. Econ.*, Oct. 1975, *1*, 423–42.

J. H. G. Olivera, "Money, Prices and Fiscal Lags: A Note on the Dynamics of Inflation," *Banca Naz. Lavoro Quart. Rev.*, Sept. 1967, *82*, 258–67.

J. D. Sargan, "Some Discrete Approximations to Continuous Time Stochastic Models," *J. Royal Statis. Soc.*, Jan. 1974, *36*, 74–90.

T. J. Sargent, and N. Wallace, "Rational Expectations and the Dynamics of Hyperinflation," *Int. Econ. Rev.*, June 1973, *2*, 328–50.

R. M. Sundrum, "Money Supply and Prices in Indonesia: 1960–70," mimeo., Australian National Univ., Dec. 1972.

———, "Money Supply and Prices: A Reinterpretation," *Bull. Indonesian Stud.*, Nov. 1973, *9*, 1–26.

R. C. Vogel, "The Dynamics of Inflation in Latin America, 1950–1969," *Amer. Econ. Rev.*, Mar. 1974, *64*, 102–14.

C. R. Wymer, "Econometric Estimation of Stochastic Differential Equation Systems," *Econometrica*, May 1972, *40*, 565–78.

———, "Continuous Time Models in Macro-Economics: Specification and Estimation," unpub. paper, 1976.

Bank of Indonesia, *Annual Report*, 1953–65.

———, *Indonesian Financial Statistics*, various issues, 1967–74.

World Bank, *Indonesia Report*, Washington 1972.

The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates

By MICHAEL DENNY AND MELVYN FUSS*

Empirical analysis of production functions has typically been pursued by postulating a substitutable relationship between aggregate indices of heterogeneous capital and labor inputs. More recently, Ernst Berndt and David Wood, and Fuss, among others, have included aggregate indices of similarly heterogeneous energy and materials inputs among the postulated factors of production. The use of aggregate input indices requires the assumption that the production function is separable in these aggregates.[1] Separability implies that marginal rates of substitution between pairs of factors in the separated group are independent of the levels of factors outside that group. Berndt and Laurits Christensen (1973b) have shown that an alternative definition is that Allen partial elasticities of substitution between a factor in the separable group and some factor outside the group be equal for all factors in the group. The separability specification substantially restricts the structure of technology and therefore the possible functional form of the production function. On the other hand, separability permits the use of aggregate data when disaggregated data are unavailable or of poor quality. Separability is consistent with decentralization in decision making, or equivalently, optimization by stages. It opens up the possibility of multistage estimation of production decisions using consistent aggregates in the latter stages. Even when adequate disaggregated data are available multistage estimation may be the only feasible procedure when large numbers of inputs are involved.[2]

Thus separability is a pivotal concept in production function estimation. Yet until recently separability and the existence of aggregate inputs were assumed a priori in virtually all production function studies. In two recent papers Berndt and Christensen (1973a, 1974) have provided the first empirical tests of separability and the possible existence of consistent aggregates of labor and capital, using one of the currently available flexible quadratic functional forms, the translog function. In carrying out their tests, Berndt and Christensen implicitly assume that the *true* underlying production function is translog (i.e., the translog function "exactly" represents the underlying production process). An alternative, more general interpretation of quadratic functional forms is that they are second-order "approximations" to some unknown arbitrary production function. This approach has been advocated by Lawrence Lau and Christensen, Dale Jorgenson, and Lau (1973, 1975). The distinction is important since the two approaches lead to different separability tests with different characteristics, even when the analysis is organized around the same functional form.

In this paper, we demonstrate first that the tests performed by Berndt and Christensen based on an exact interpretation of the translog function are more restrictive than is readily apparent and therefore cannot be accepted as general tests of the separability

*Associate professors of economics, University of Toronto; research associates, Institute for Policy Analysis. We wish to acknowledge helpful comments from Ernst Berndt, Lawrence Lau, and an anonymous referee, but retain responsibility for any remaining errors. This paper is a condensed and somewhat simplified version of our working paper with the same title, to which the reader is referred for technical elaboration of the concepts presented.

[1]We are implicitly assuming that factor prices do not vary in proportion, so that Hicks' aggregation theorem does not provide a means of justifying the use of aggregates.

[2]For an example see the two-stage procedure used by Fuss.

hypothesis.[3] In particular, the separable functions must be either Cobb-Douglas functions of translog subaggregates or translog functions of Cobb-Douglas subaggregates.[4] Second, we develop tests of this property based on the translog function as a quadratic approximation which are less restrictive than the Berndt-Christensen tests, since they do not impose a Cobb-Douglas structure on any portion of the underlying function.[5] Finally, we apply the approximate tests to the Berndt-Christensen data. We find little evidence in support of labor aggregation, confirming the Berndt-Christensen results. Separability of equipment and structures from aggregate labor cannot be rejected, but the form of the separability is not that accepted by Berndt-Christensen. Somewhat to our surprise, we find that after estimating a "flexible" functional form—the translog—we find evidence of the existence of complete strong separability in the form of a "rigid" functional form—the three-factor constant elasticity of substitution (*CES*) function. This structure is strongly supported by evidence from the cost function estimates. The production function results are less definitive and could reasonably encompass the alternative interpretation.

I. The Exact Test for Separability

The basic problem with the Berndt-Christensen exact tests for separability is that they are not just tests of the null hypothesis of separability. Rather, they are tests of the *joint* null hypothesis of separability and a particular ***inflexible*** functional form for either the aggregator functions or the production function as a function of aggregate inputs.[6] This result is contained in Proposition 1.

PROPOSITION 1: *The separable form of a translog function interpreted as an exact production function must be either a Cobb-Douglas function of translog subaggregates or a translog function of Cobb-Douglas subaggregates.*

Since our empirical analysis involves the uses of three inputs we will demonstrate this assertion for the three-input case. This is the example used by Berndt and Christensen (1973a, 1974). In this section we provide an intuitive derivation of our assertion. A more formal proof for the three-input case is outlined in the Appendix. The proof of this result for an arbitrary number of factors partitioned into two aggregates or for an arbitrary number of aggregates involves a straightforward extension of the elements of the proof contained in the Appendix. For this extension see Blackorby, Primont, and Russell. Let the true production function be $\ln y = f(\ln X_1, \ln X_2, \ln X_3)$ where y is output and $X_i, i = 1, 2, 3$ are inputs.[7] If we impose the condition that f is translog in all three inputs then it must be a quadratic function in those inputs and take the form

$$(1) \quad \ln y = \ln \alpha_0 + \sum_{i=1}^{3} \alpha_i \ln X_i + \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \gamma_{ij} \ln X_i \ln X_j$$

Suppose now the true production function is weakly separable in X_1 and X_2; i.e.,

[3]At the time this paper was being written we became are of a paper by Charles Blackorby, Daniel imont, and Robert Russell which also contains this monstration.

[4]The other flexible functional form which appears in : literature, W.E. Diewert's generalized Leontief, o has a restrictive separable form when viewed as act The linear separability conditions imply that the oduction function is a linear function of generalized ontief aggregates. The non-linear separability condi- ns imply that the function is a generalized Leontief nction of *CES* aggregates. Within each of the aggre- tes the subaggregates have a common elasticity of bstitution equal to 2.

[5]These tests are similar in nature to the approximate sts found in Christensen, Jorgenson, and Lau (1975) id some of them are derived in Jorgenson and Lau r the case of translog utility functions. For example, roposition 4 below is called groupwise separability v Jorgenson and Lau. Proposition 5 is called in- insic/explicit groupwise separability.

[6]By inflexible we mean that once separability has been imposed, either the aggregator functions or the function of aggregates is no longer capable of providing an arbitrary second order approximation to the separable technology.

[7]The function is written in terms of logarithms without loss of generality since $y = g(X_1 \ldots X_N)$ implies

$$\ln y = \ln g(e^{\ln x_1} \ldots e^{\ln x_N}) = f(\ln X_1 \ldots \ln X_N)$$

(2) $\ln y = f[\ln G(\ln X_1, \ln X_2), \ln X_3]$

where G is an input aggregator function. Clearly there are only two possibilities:

(i) $\ln G$ is quadratic in $\ln X_1$ and $\ln X_2$. Then $f(\ln G, \ln X_3)$ can be only linear in $\ln G$ and $\ln X_3$. In the Appendix we show that this case corresponds to the constraints $\gamma_{13} = \gamma_{23} = 0$, which are Berndt and Christensen's linear separability constraints. Equation (2) takes the form

(3) $\ln y = \ln \alpha_0 + \theta_G \ln G + \theta_H \ln H$

where θ_G, θ_H are parameters, G is a translog function of X_1 and X_2, and H is a translog function of X_3. The production function is a Cobb-Douglas function of translog aggregates.

(ii) $\ln G$ is linear in $\ln X_1$ and $\ln X_2$. Then f can only be quadratic in $\ln G$ and $\ln X_3$. In the Appendix we show that this case corresponds to the constraints $\alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23} = \gamma_{11}/\gamma_{21} = \gamma_{12}/\gamma_{22}$, which are Berndt and Christensen's non-linear separability constraints. Equation (2) takes the form

(4) $$\ln y = \ln \alpha_0 + \beta_G \ln G + \beta_H \ln H + \frac{1}{2}\sum_{i,j=G,H}\sum \beta_{ij} \ln G \ln H$$

where $\ln \alpha_0$, β_G, β_H, β_{ij} are parameters. The non-linear separability restrictions imply that the aggregate input G is a Cobb-Douglas combination of components X_1 and X_2. The production function is translog in the aggregates. Therefore non-linear translog separability implies a unitary subaggregate elasticity of substitution between X_1 and X_2 (i.e., along a G isoquant) as well as $\sigma_{13} = \sigma_{23}$ (where σ_{ij} is the Allen partial elasticity of substitution between factors i and j along a y isoquant). Rejection of the separability hypothesis may be the result of rejecting a subaggregate elasticity of unity instead of the correct null hypothesis $\sigma_{13} = \sigma_{23}$. In order to avoid this problem we need to consider the translog function not as an exact function but rather as a quadratic approximation.

II. Approximation Analysis and Tests of Hypotheses

In this section we describe procedure for using approximation analysis in the testing of hypotheses, and develop separability tests based on these procedures. We confine ourselves to a quadratic approximation. The extension to higher order approximations should be readily apparent.

DEFINITION 1: A second-order *approximation* to the production function $Q = f(z)$, where $z = [Z_1 \ldots Z_N]$, is the Taylor series quadratic expansion

(5) $$\hat{Q}(z) = f(z^*) + \sum_{i=1}^{N} \frac{\partial f}{\partial Z_i}\bigg|_{z^*} \cdot [Z_i - Z_i^*] + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2 f}{\partial Z_i \partial Z_j}\bigg|_{z^*} \cdot [Z_i - Z_i^*][Z_j - Z_j^*]$$

where $z^* = [Z_1^* \ldots Z_N^*]$ is the point of expansion.

DEFINITION 2: The expansion $\hat{Q}(Z)$ is *exact* for Q if $Q = \hat{Q}(z)$ for all z.[8]

The following definition provides the key to our development of approximate hypothesis tests.

DEFINITION 3: Two production functions $Q_1 = f_1(z)$ and $Q_2 = f_2(z)$ represent the same underlying technology *up to a second-order approximation* if their quadratic expansions $\hat{Q}_1(z)$, $\hat{Q}_2(z)$ around any common expansion point z^* are identical for all input vectors z.

Since two expansions of the form (5) will differ only in their parameters $f(z^*)$, $(\partial f/\partial Z_i)|_{z^*}$, $(\partial^2 f/\partial Z_i \partial Z_j)|_{z^*}$, Definition 3 has the following equivalent alternative interpretation: two production functions represent the same underlying technology

[8]This definition of exactness, i.e., that a particular condition should hold for every value of the vector, is used in the Appendix to label the Berndt-Christensen separability restrictions as exact restrictions.

(approximately) if the parameters of their expansions are identical.

We are now in a position to consider the translog function as a quadratic approximation.

PROPOSITION 2: *The translog function of the form* (1) *with symmetry imposed* ($\gamma_{ij} = \gamma_{ji}$) *is a quadratic approximation (around the expansion point* $x^* = [X_1^* \ldots X_N^*] = [1, \ldots 1]$) *to an arbitrary production function of the form* $\ln y = f[\ln X_1, \ldots \ln X_N]$.

PROOF:

Define $Q[Z] = \ln y$; $Z_i = \ln X_i$; $\alpha_0 = f(z^*)$; $\alpha_i = \left.\frac{\partial f}{\partial Z_i}\right|_{z^*}$

$$\gamma_{ij} = \left.\frac{\partial^2 f}{\partial Z_i \partial Z_j}\right|_{z^*} = \left.\frac{\partial^2 f}{\partial Z_j \partial Z_i}\right|_{z^*} = \gamma_{ji}$$

Also $Z_i^* = \ln 1 = 0$. Then

$$(6)\quad \hat{Q} = \ln \hat{y} = \alpha_0 + \sum_{i=1}^{N} \alpha_i \ln X_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \gamma_{ij} \ln X_i \ln X_j$$

ith $\gamma_{ij} = \gamma_{ji}$ satisfies Definition 1.[9]

ROPOSITION 3: *The translog function of* *te form* (1) *with symmetry* ($\gamma_{ij} = \gamma_{ji}$) and the *dding up conditions* ($\sum_i \alpha_i = 1$, $\sum_j \gamma_{ij} = 0$) *nposed is a quadratic approximation (around he expansion point* $x^* = 1$) *to an arbitrary near homogeneous production function.*

'ROOF:

Consider the Euler equations for linear iomogeneity of an arbitrary production unction $y = g(X_1 \ldots X_N)$. These conditions ire

$$(7)\quad \sum_{i=1}^{N} \frac{\partial \ln y}{\partial \ln X_i} = 1$$

and

$$\sum_{j=1}^{N} \frac{\partial^2 \ln y}{\partial \ln X_i \partial \ln X_j} = 0 \qquad i = 1, \ldots N$$

Using the same definitions as in Proposition 2, constraints (7) can be written

$$(8)\quad \sum_{i=1}^{N} \frac{\partial f}{\partial Z_i} = 1$$

and

$$\sum_{j=1}^{N} \frac{\partial^2 f}{\partial Z_i \partial Z_j} = 0$$

Therefore equation (5) subject to (8) is a quadratic approximation which satisfies the linear homogeneity requirements. Now consider the translog approximation (6):

$$(9)\quad \sum_{i=1}^{N} \left.\frac{\partial \ln \hat{y}}{\partial \ln X_i}\right|_{x^*} = \sum_i \alpha_i$$

and

$$\sum_{j=1}^{N} \left.\frac{\partial^2 \ln \hat{y}}{\partial \ln X_i \partial \ln X_j}\right|_{x^*} = \sum_j \gamma_{ij}$$

When $\sum_i \alpha_i = 1$ and $\sum_j \gamma_{ij} = 0$, equation (6) subject to (9) is identical to equation (5) subject to (8). Therefore Definition 3 is satisfied and the proposition is proved.

Condition (9) is normally imposed in the estimation procedure along with the symmetry conditions. Therefore linear homogeneity will be a maintained hypothesis in the empirical portion of this paper.[10]

A. *Approximate Separability Conditions*

In this section we derive our approximate tests for weak and strong separability. We

[9] The symmetry constraints $\gamma_{ij} = \gamma_{ji}$ are imposed by he equality of cross-partial derivatives in a quadratic xpansion. The exact interpretation of the translog unction does not contain this implication. In that ase, symmetry is usually imposed to obtain identificaion of the parameters since only $\gamma_{ij} + \gamma_{ji}$ is identified nitially.

[10] Constraints (9) are also imposed when the translog is assumed to be an exact representation. It is easily seen that they also can be used to satisfy Definition 2 since

$$\sum_j \frac{\partial^2 \ln y}{\partial \ln X_i \partial \ln X_j} = \sum_j \gamma_{ij} = 0$$

implies $\sum_i \gamma_{ij} = 0$ from symmetry; and, therefore,

$$\sum \frac{\partial \ln y}{\partial \ln X_i} = \sum_i \alpha_i + \sum_i \gamma_{ij} \ln X_j = \sum \alpha_i$$

TABLE 1—SPECIFICATION OF SEPARABILITY RESTRICTIONS FOR THE FUNCTION $f[\ln X_1, \ln X_2, \ln X_3]$

Hypothesis	Functional Form	Proposition
Weak separability	$f[G(\ln X_1, \ln X_2), \ln X_3]$	4
Logarithmic partial strong separability	$G(\ln X_1, \ln X_2) + H(\ln X_3)$	5
Complete strong separability (with constant elasticity of substitution)	$\gamma[\delta_1 X_1^\rho + \delta_2 X_2^\rho + \delta_3 X_3^\rho]^{1/\rho}$	6
Logarithmic complete strong separability	$\alpha_0 + \alpha_1 \ln X_1 + \alpha_2 \ln X_2 + \alpha_3 \ln X_3$	7

have used this division instead of the Berndt-Christensen linear and non-linear restrictions dichotomy because of its more fundamental economic meaning. The two classifications are related, as is demonstrated below.

We now return to our three-input case for ease of exposition. Extensions to more than three factors will be indicated where appropriate. Propositions 4–7 below present the separability restrictions tested in this paper. Table 1 summarizes these restrictions in general functional notation for the case of X_1, X_2 separable from X_3.[11]

PROPOSITION 4: *The translog function* (1) *is a quadratic approximation to an arbitrary weakly separable production function*

$$\ln y = f[G(\ln X_1, \ln X_2), \ln X_3] \tag{10}$$

$$\text{if } \alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23} \tag{11}$$

The proof of Proposition 4 consists of expanding (10) as a quadratic expansion around $x^* = 1$ and finding the restriction on (6) such that Definition 3 is satisfied. The details are provided in the Appendix.

Comparing constraint (11) with the conditions for separability in the exact case $(\alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23} = \gamma_{11}/\gamma_{21} = \gamma_{12}/\gamma_{22})$ we see that (11) is identical to the first set of constraints. For the situation of inputs 1 and 2 separable from 3, Berndt and Christensen (1973a) have shown that the remainder of the constraints reduce to 1 independent constraint of the form $\gamma_{11} \cdot \gamma_{22} = (\gamma_{12})^2$. But this constraint is just the condition which forces $G(\ln X_1, \ln X_2)$ to be a Cobb-Douglas aggregate, as can be seen from the proof of Proposition 1. To test approximate weak separability we test the null hypothesis $\alpha_1\gamma_{23} - \alpha_2\gamma_{13} = 0$. The hypothesis of exact weak separability is nested in the approximate test since it involves the constraint $\gamma_{11} \cdot \gamma_{22} - (\gamma_{12})^2 = 0$ as well as the constraint $\alpha_1\gamma_{23} - \alpha_2\gamma_{13} = 0$.[12] Therefore a natural sequence of hypotheses is established moving from the maintained hypothesis which does not impose separability to approximate weak separability to exact weak separability.

Weak separability is a necessary but not sufficient condition for the two-stage optimization procedure implicit in the existence of consistent aggregates. A sufficient condition is weak homothetic separability. However, when the function is assumed to be linear homogeneous, the conditions for weak separability and weak homothetic separability are the same.[13] This result is stated as Proposition 4A and proved in the Appendix.

PROPOSITION 4A: *Weak separability of a linear homogeneous function is equivalent to weak homothetic separability of the function.*

[11] We replace $\ln G$ by G for ease of notation without loss of generality.

[12] More generally, consider any partition of the factors where X_i, X_j are in the partition and X_k is excluded. Approximate weak separability involves imposing the constraints $\alpha_i\gamma_{jk} - \alpha_j\gamma_{ik} = 0$. Exact weak separability requires the imposition of the additional constraints $\gamma_{ii} \cdot \gamma_{jj} - (\gamma_{ij})^2 = 0$.

[13] A function is said to be weakly homothetic separable if it is weakly separable and the subaggregates are homothetic in their arguments.

Proposition 4A implies that the tests for weak separability are also tests for weak homothetic separability and can be viewed as tests of the existence of consistent aggregates within the approximation framework. Since strong separability is a special case of weak separability, this result also holds for the strong separability tests to be introduced in the next three propositions.

PROPOSITION 5: *The translog function is a quadratic approximation to an arbitrary logarithmic partially strong separable production function of the form*

$$(12)\quad \ln y = G(\ln X_1, \ln X_2) + H(\ln X_3)$$

$$(13)\qquad \text{if } \gamma_{13} = \gamma_{23} = 0$$

The proof of Proposition 5 is obtained in the same manner as that of Proposition 4; i.e., by expanding (12) as a quadratic approximation around $x^* = 1$ and comparing the result with equation (6) to obtain the constraints (13). The quadratic approximation to the more general form of partial strong separability,

$$(14)\quad \ln y = f[G(\ln X_1, \ln X_2) + H(\ln X_3)]$$

is identical to the weak separability approximation when there are only two partitions and so is not pursued in this paper. When there are three or more partitions approximate partial strong separability is more restrictive.[14]

The strong separability constraints (13) are identical to the exact linear constraints. For this case of strong separability, the exact and approximate interpretations of the translog function yield the same testable hypotheses. Hence, using the results of Section I, separability is of the Cobb-Douglas or logarithmic type. However, unlike the two exact tests, the approximate tests do not yield conflicting criteria with respect to weak and strong separability.[15] Since $\gamma_{13} = \gamma_{23} = 0$ is a sufficient but not necessary condition for approximate weak separability, the approximate test for strong separability is nested within the one for weak separability. This is a desirable feature since strong separability is more restrictive than weak separability.

The next proposition which we present in this paper relates to the use of the translog function as an approximation to the *CES* function. If X_1 and X_2 are weakly separable from X_3; and simultaneously X_1 and X_3 are weakly separable from X_2 then $\sigma_{13} = \sigma_{23}$ and $\sigma_{12} = \sigma_{32}$ $(=\sigma_{23})$. The pairwise equality of these Allen partial elasticities of substitution implies $\sigma_{13} = \sigma_{23} = \sigma_{12}$, or equality of all elasticities of substitution. Berndt and Christensen (1973b) have shown that equality of all σ_{ij} implies a completely strongly separable function which can be written in the form

$$(14)\quad \ln y = f[G_1(\ln X_1) + G_2(\ln X_2) + G_3(\ln X_3)]$$

Given our assumption of linear homogeneity, Proposition 4A implies that the G_i are homothetic. Therefore imposition of the constraints $\alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23}$ and $\alpha_1/\alpha_3 = \gamma_{12}/\gamma_{23}$ yields a translog approximation to a homothetically strong separable function (14).

If the σ_{ij} are constant, Hirofumi Uzawa has shown that the underlying function (14) is three-factor *CES* of the form

$$(15)\quad y = \gamma[\delta_1 X_1^{\rho} + \delta_2 X_2^{\rho} + \delta_3 X_3^{\rho}]^{1/\rho}$$

Since the elasticities of substitution for the translog function will be constant when evaluated at a given point (in this case x^*), simultaneous imposition of two approximate weak separability constraints yields an approximation to the *CES* function.[16] This is the substance of Proposition 6.

[14]Consider the case of 4 inputs and 3 partitions. Weak separability of the form $\ln y = f[G(\ln X_1, \ln X_2), \ln X_3, \ln X_4]$ can be approximated by the translog function with constraints $\alpha_1\gamma_{23} = \alpha_2\gamma_{13}$ and $\alpha_1\gamma_{24} = \alpha_2\gamma_{14}$ using the techniques of Proposition 4. Partial strong separability of the form $\ln y = f[G(\ln X_1, \ln X_2) + H(\ln X_3) + J(\ln X_4)]$ can be approximated by adding the constraints $\alpha_2\gamma_{34} = \alpha_3\gamma_{24}$ and $\alpha_1\gamma_{34} = \alpha_4\gamma_{13}$. This can be proved by quadratic expansion and use of Definition 3.

[15]For the exact tests, $\gamma_{13} = \gamma_{23} = 0$ satisfies both the strong and weak separability requirements but the additional weak separability constraint $\gamma_{11} \cdot \gamma_{22} = (\gamma_{12})^2$ is not required for strong separability thus causing nonnested potential conflicts.

[16]For a k factor production function, simultaneous imposition of $(k/2) - 1$ approximate weak separability conditions is sufficient to obtain pairwise equality of all partial elasticities of substitution.

PROPOSITION 6: *The linear homogeneous translog function is a quadratic approximation to an arbitrary CES production function of the form* (15) *if:* (a) *the adding up constraints* $\sum_i \alpha_i = 1$, $\sum_j \gamma_{ij} = 0$ *and* (b) *the approximate weak separability constraints* $\alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23}$ $(\sigma_{13} = \sigma_{23})$ *and* $\alpha_1/\alpha_3 = \gamma_{12}/\gamma_{23}$ $(\sigma_{12} = \sigma_{32})$ *hold simultaneously.*[17]

The proof of this proposition is identical to those of Propositions 4 and 5. The reader is referred to the authors' working paper for a detailed proof.

Berndt and Christensen (1973b) have shown that, given constant returns to scale, equation (15) is the most general form of complete strong separability. Therefore the constraints of Proposition 5 provide the specification for approximate complete strong separability. This specification is nested in the approximate weak separability specification. The technique of sequential testing is again available.

It is well known that the Cobb-Douglas production function is a special case of the *CES* function (15) with $\rho = 0$. Our quadratic approximation has the same property and this fact leads to the parameter restrictions of Proposition 7.

PROPOSITION 7: *The linear homogeneous translog function is a quadratic approximation to an arbitrary Cobb-Douglas production function of the form*

$$(16) \quad \ln y = \alpha_0 + \alpha_1 \ln X_1 + \alpha_2 \ln X_2 + \alpha_3 \ln X_3$$

if, in addition to the constraints of Proposition 6, the constraints $\gamma_{12} = \gamma_{13} = 0$ *hold.*

The proof of this proposition is entirely analogous to those of the earlier propositions. For details see the authors' working paper. Note that the hypothesis of Cobb-Douglas specification is nested in the *CES* specification (in terms of our approximate tests). Thus a sequential testing procedure is available in this case as well.

[17]Note that these constraints imply the third weak separability constraint $\alpha_2/\alpha_3 = \gamma_{12}/\gamma_{13}$.

III. Empirical Results

In order to implement the above hypothesis testing procedure, we have estimated translog production and cost structures[18] for U.S. Manufacturing 1929-68. We use the Berndt-Christensen data, which consist of observations for blue collar workers (B), white collar workers (W), and capital (K) in one case (1974); and on equipment (E), structures (S), and labor (L) in the other case (1973a). The reader is referred to the cited papers for details as to the construction of these data.

A. *Estimation of the Model under the Maintained Hypothesis*

Berndt and Christensen (1973a) have shown that the linear homogeneous production structure can be obtained by estimating the system of demand equations

$$(17) \quad S_i = \alpha_i + \sum_{j=1}^{3} \gamma_{ij} \ln X_j \qquad i = 1,2,3$$

subject to $\sum \alpha_i = 1$, $\gamma_{ij} = \gamma_{ji}$, $\sum_j \gamma_{ij} = ($ where S_i is the cost share of the ith facto Similarly it can be shown (see Berndt an Wood) that a linear homogeneous co structure can be estimated from the syste of demand functions

$$(18) \quad S_i = \alpha_i + \sum_{j=1}^{3} \gamma_{ij} \ln p_j \qquad i = 1,2,$$

subject to $\sum \alpha_i = 1$, $\gamma_{ij} = \gamma_{ji}$, $\sum_j \gamma_{ij} =$ where S_i is the cost share as before and p_j i the price of the jth factor.

Table 2 presents our parameter estimate and a comparison with the Berndt-Christen sen results for the production structures (They did not estimate the cost structures. There are two differences between their esti mation procedure and ours. First, we d not employ instrumental variables for th X_i, thus ignoring possible simultaneity bias Second, we employ two-stage Zellner effi

[18]The translog cost function with linear homo geneity imposed can be written $\ln C = \alpha_0 + \sum_i \alpha_i \ln p_i + 1/2 \sum_i \sum_j \gamma_{ij} \ln p_i \ln p_j$. All of the separabilit results of the previous section hold for the cost func tion with X_i replaced by p_i.

TABLE 2—PARAMETER ESTIMATES FOR NONSEPARABLE TRANSLOG FUNCTIONS (MAINTAINED HYPOTHESIS)[a]

	Two Types of Capital-One Type of Labor: X_1 equipment, X_2 = structures, X_3 = labor		Two Types of Labor-One Type of Capital: X_1 = blue collar, X_2 = white collar, X_3 = capital	
Parameter	Production This Study (1)	Production Berndt-Christensen (2)	Production This Study (3)	Production Berndt-Christensen (4)
α_1	0.0979	0.0990	0.6260	0.6274
	(0.0026)	(0.0027)	(0.0033)	(0.0042)
α_2	0.0742	0.0734	0.2159	0.2130
	(0.0023)	(0.0023)	(0.0034)	(0.0039)
α_3	0.8279	0.8276	0.1581	0.1595
	(0.0033)	(0.0034)	(0.0033)	(0.0037)
γ_{11}	0.0280	0.0335	0.1793	0.1833
	(0.0096)	(0.0102)	(0.0095)	(0.0129)
γ_{22}	0.0355	0.0318	0.0958	0.1107
	(0.0064)	(0.0065)	(0.0127)	(0.0145)
γ_{33}	0.0324	0.0399	−0.0028	0.0014
	(0.0146)	(0.0155)	(0.0123)	(0.0137)
γ_{12}	−0.0155	−0 0127	−0.1389	−0.1463
	(0.0053)	(0.0055)	(0.0076)	(0.0092)
γ_{13}	−0.0125	−0.0208	−0.0404	−0.0370
	(0.0106)	(0.0113)	(0.0080)	(0.0097)
γ_{23}	−0.0200	−0.0191	0.0431	0.0357
	(0.0077)	(0.0079)	(0.0108)	(0.0123)

[a]Standard errors are in parentheses.

cient estimation, choosing not to iterate on our initial solution. A comparison of columns (1, 2) and (3, 4) in Table 2 indicates that the different estimation procedures result in roughly identical structures. Therefore the interested reader may compare our approximate tests with the Berndt-Christensen exact tests.

B. *Tests of the Separability Hypotheses*

Tables 4–5 present the results of applying r sequential tests for separability for both production and cost structures to the o capital-one labor and two labor-one pital cases, respectively. Two aspects of e theory of sequential testing are emoyed. First, the sequence ends when the st null hypothesis which can be rejected is countered. Second, the appropriate significance level of the test depends on the gnificance levels of the prior hypotheses in e sequence which could not be rejected. ippose H_0^i is the ith null hypothesis in a particular sequence with significance level δ, conditional on sequential nonrejection of the prior hypotheses. Then the appropriate significance level for testing H_0^i is[19]

$$1 - \prod_{j=1}^{i} (1 - \delta_j) \simeq \sum_{j=1}^{i} \delta_j$$

For the purposes of this paper we have chosen $\delta_1 = 0.01, \delta_2 = 0.015, \delta_3 = 0.025$. We use the approximate F-statistic for our tests. Appropriate significance levels, degrees of freedom, and critical values for the various tests in the sequence are contained in Table 3.

We first consider the two common separability assumptions: (E, S), L and (B, W), K. Using the significance levels sequence: 1 percent, 2.5 percent, 5 percent, we cannot reject the hypothesis of weak separability of equipment and structures from aggregate labor, either on the production or cost side.

[19]For an accessible explanation of sequential nested hypothesis testing see Edmund Malinvaud, pp. 218–20.

TABLE 3—SIGNIFICANCE LEVELS AND CRITICAL VALUES USED FOR SEQUENTIAL TESTING

Test	Significance Level	Degrees of Freedom	Critical Values
Weak Separability (I)	.01	(1,75)	7.02
Logarithmic partial strong separability (IIa)	.025	(1,76)	5 25
Complete strong separability (IIb)	.025	(1,76)	5.25
Logarithmic complete strong separability (III)	.05	(1,77)	3 98

We reject logarithmic partial strong separability.[20] Complete strong separability is rejected on the production side but not on the cost side. There is strong evidence for the existence of a *CES* cost function. Since the *CES* function is self-dual,[21] the resultant production and cost structures are potentially conflicting descriptions of technology. This issue is pursued below.

We reject the hypothesis that the two types of labor are separable from aggregate capital both on the production side and the cost side. On the basis of these results it would appear that neither the existence of labor aggregation nor the existence of an aggregate price index for labor can be supported.

We now turn to the overall evidence of separability, regardless of whether the type of separability is conventional. For the *E*, *S*, *L* case the evidence throughout the rest of Table 4 is consistent with the three-factor *CES* production and cost functions specification.[22] There are twelve null hypothes related to the *CES* structure (I and IIb f each of the six rows in the Table). Rejecti of any one of these is logically a rejection the possibility of the existence of an overa *CES* structure. Only one of the twelve w rejected at the significance levels chosen, a though several of the weak separability tes result in marginal nonrejection. There a pears to be some overall evidence that in position of a *CES* structure is roughly co sistent with the underlying data. For mar applications the simplicity thus obtaine probably outweighs the bias introduced.

For the *B*, *W*, *K* case, the results are mo mixed. All forms of weak separability o the production side are decisively rejecte There appears to be evidence of parti strong logarithmic separability on the co side of the form *B*, (*W*, *K*). This sugges there exists an aggregate price index fo white collar workers and capital which i compatible with the data.[23]

We now turn to a consideration of th structure of technology as estimated by ou nonrejected functional forms. Final param eter estimates are contained in Table 6 Partial elasticities of substitution for se lected years are presented in Table 7.[24] W concentrate on an analysis of the substitu tion elasticities since the parameters b themselves are difficult to interpret for flex ible functional forms such as the translog.

For the *E*, *S*, *L* case we will consider th completely strong separable structure. Th implies that we are approximating constan elasticity of substitution functions with con

[20]This is the form of separability that was not rejected by Berndt and Christensen (1973a). We also would not reject it using a simultaneous testing procedure at a 1 percent significance level rather than a sequential one, since our test statistic would have been 3.45. However, as we demonstrate below, once the cost structure is included in the analysis, nonlogarithmic separability is probably more consistent with the underlying data.

[21]Care must be taken in interpreting this duality. The translog approximation to a *CES* production (cost) function is not itself self-dual (except at the point of approximation). Self-duality is a consequence of not rejecting *CES* functions on both the production and cost sides, and as a result imposing these structures.

[22]At significance levels much greater than that used in this paper, there is evidence that the *CES* production structure would be rejected within all three sequences, but not the cost structure. The strong evidence on the cost side suggests that if we wish to assume that both the cost and production functions represent the same underlying technology we should not reject complete strong separability on the production side.

[23]This case appears to be one where the Hicks aggregation theorem is applicable. The quantity indices for white collar workers and capital are highly correlated with $\rho_{WK} = .92$, whereas $\rho_{BW} = .77$, and $\rho_{BK} = .65$

[24]The final parameter estimates for the *BWK* production case can be found in Table 4 since all separability restrictions were rejected.

TABLE 4—SEPARABILITY TEST STATISTICS: EQUIPMENT (E), STRUCTURES (S), LABOR (L)

		Sequences of Hypotheses[a]				
					III	
Partition	Structure	I	IIa	IIb	Conditional on IIa	Conditional on IIb
$(E, S), L$	Production	1.22	5.60	5.99		33.34
	Cost	1.37	5.65	0.01		7.78
$S, (E, L)$	Production	4.95	28.94	2.21		33.34
	Cost	0.01	0.52	1.37	8.73	7.78
$E, (S, L)$	Production	6.11	2.46	1.11	31.39	33.34
	Cost	0.02	8.10	1.36		7.78

[a]Refer to Table 3 for a description of these hypotheses.

TABLE 5—SEPARABILITY TEST STATISTICS: BLUE COLLAR (B), WHITE COLLAR (W), CAPITAL (K)

		Sequences of Hypotheses[a]				
					III	
Partition	Structure	I	IIa	IIb	Conditional on IIa	Conditional on IIb
$(B, W), K$	Production	17.70				
	Cost	9.01				
$B, (W, K)$	Production	19.51				
	Cost	3.08	3.38	31.78	28.42	
$(B, K), W$	Production	39.34				
	Cost	12.13				

[a]Refer to Table 3 for a description of these hypotheses

TABLE 6—PARAMETER ESTIMATES FOR SEPARABLE TRANSLOG FUNCTIONS[a]

	(1) Two Types of Labor – One Type of Capital: X_1 = Blue Collar; X_2 = White Collar; X_3 = Capital	(2) Two Types of Capital – One Type of Labor: X_1 = Equipment; X_2 = Structures; X_3 = Labor	(3)
Parameter	Cost Logarithmic Partial Strong Separability	Production *CES*	Cost *CES*
α_1	0.5914 (0.0066)	0.1001 (0.0023)	0.0858 (0.0035)
α_2	0.2437 (0.0047)	0.0762 (0.0021)	0.0766 (0.0029)
α_3	0.1650 (0.0028)	0.8238 (0.0029)	0.8376 (0.0035)
γ_{11}	0.0000	0.0373 (0.0063)	0.0198 (0.0068)
γ_{22}	0.0408 (0.0060)	0.0291 (0.0047)	0.0178 (0.0069)
γ_{33}	0.0408 (0.0060)	0.0601 (0.0099)	0.0343 (0.0125)
γ_{12}	0.0000	−0.0032 (0.0005)	−0.0017 (0.0006)
γ_{13}	0.0000	−0.0341 (0.0058)	−0.0181 (0.0062)
γ_{23}	−0.0408 (0.0060)	−0.0260 (0.0042)	−0.0162 (0.0063)

[a]Standard errors are in parentheses.

TABLE 7—ALLEN PARTIAL ELASTICITIES OF SUBSTITUTION-SELECTED YEARS
(corresponding to final technology structures)

	Separability Structure	1929	1939	1949	1959	1968
ESL						
Production						
σ_{ES}		1.836	1.829	1.706	1.873	1.919
σ_{EL}	IIb	1.919	1.915	1.706	1.694	1.649
σ_{SL}		1.616	1.614	1.706	1.857	1.962
Cost						
σ_{ES}		0.777	0.769	0.748	0.780	0.768
σ_{EL}	IIb	0.768	0.774	0.748	0.768	0.764
σ_{SL}		0.752	0.737	0.748	0.754	0.748
BWK						
Production						
σ_{BW}		5.251	5.633	6.990	4.896	4.763
σ_{BK}	No separability	3.018	3.160	3.696	2.900	2.859
σ_{WK}		−2.646	−3.216	−5.186	−2.103	−1.882
Cost						
σ_{BW}		1.000	1.000	1.000	1.000	1.000
σ_{BK}	IIa	1.000	1.000	1.000	1.000	1.000
σ_{WK}		−0.003	0.001	−0.014	−0.002	−0.017

stancy imposed at the point of expansion. All quantities and prices were normalized so that the point of expansion $x^* = (1, 1, 1)$ occurred in the year 1949. From Table 7 it can be seen that the partial elasticities of substitution among inputs are indeed equal in 1949 at 1.706 for the production structure and 0.748 for the cost structure. The small variation, both across factors and across time, indicates that once complete strong separability is imposed, the second-order approximation does in fact closely resemble a *CES* function for this data set. If the cost and production functions are to be self-dual, the elasticities of substitution obtained from these functions should be the same. The point estimates are quite different, but they are likely to be inaccurately estimated. These elasticities are complex functions of the parameter estimates but a rough check can be made using confidence intervals based on linearized approximate standard deviations calculated at the point of expansion. A .01 approximate confidence interval for σ_{ij} is (2.19, 1.32) and (.94, .56) for the production and cost functions, respectively. The actual distributions of the estimated elasticities will not be the same as these (assumed) normally distributed linearized approximations. Therefore, it is difficult to draw conclusions. However, the fact that these intervals do not overlap implies that the estimated cost and production structures may be approximating different *CES* structures which are not self-dual.

For the *B*, *W*, *K* case, we rejected all forms of separability on the production side. The elasticities calculated from the production structure exhibit considerable variation both among factors and across time. White collar workers and aggregate capital are estimated to be complementary factors of production. On the cost side we stopped our sequential testing with the non-rejection of partial logarithmic strong separability of the form $f(B, W, K) = \ln G(B) + \ln H(W,K)$. Imposing this structure requires $\sigma_{BW} = \sigma_{BK} = 1$. This result also appears in Table 7. From the cost structure we estimate that *W* and *K* are very weak complements. A comparison of substitution elasticities as derived from the cost and production structures demonstrates that the direct and dual formulations are competing descriptions of technology. The sharply different estimates of the elasticities of substitution arise because we cannot reject partial strong logarithmic separability for the cost case. The elasticities for the *BWK* cost case with no separability imposed are similar to those

TABLE 8—E, S, L PRODUCTION STRUCTURE: E SEPARABLE FROM S AND L

	Maintained hypothesis	Weak separability	Logarithmic partial strong separability	Complete strong separability
Conditional F-statistic	—	6.110	2.460	1.100
σ_{ES}[a]	7.250	1.499	1.000	1.706
σ_{EL}[a]	1.089	1.499	1.000	1.706
α_{SL}[a]	1.623	1.955	1.984	1.706

[a]Elasticities are calculated at the point of expansion (year 1949).

or the production case with no separability mposed. For BWK cost they are σ_{BW} = '06, σ_{BK} = 1.50, and σ_{WK} = −0.91, which :an be compared with those given in Table ' for BWK production.

Finally, we illustrate the sensitivity of the)artial elasticities of substitution to the im-)osition of successive structures when some)f the reject-nonreject decisions are mar- ginal. For the ESL production function :ase we present in Table 8 partial elasticities)f substitution corresponding to the imposi- tion of the successive hypotheses.

The imposition of E, (SL) weak separa- bility forces the equality of σ_{ES} and σ_{EL}. These elasticities are quite different from those of the maintained hypothesis. When we impose logarithmic partial strong sepa- rability, these elasticities are lowered to one and σ_{SL} is forced up further. The approxi- mate CES constraints are readily accepted :e weak separability is maintained since elasticities are relatively close. The very ge substitution elasticity between equip- nt and structures is eliminated when = 1.706. These results illustrate the diffi- ties inherent in attempting to obtain pre- : estimates of elasticities of substitution.

IV. Conclusions

Like Berndt-Christensen, we find little dence to support labor separability. The le exception is the possibility of forming aggregate price index of capital and iite collar workers. But this possibility is obably a property of the particular data ed rather than of the underlying tech- ology. Also like Berndt and Christensen : find support for capital aggregation— both in terms of quantity and price in- dices—once labor aggregation is assumed a priori. However, our approximate testing procedure allows us to investigate the nature of the separability in more detail. Somewhat surprisingly, we found that the specification of complete strong separability in the form of three-factor CES cost and production functions is roughly consistent with the underlying data although formally rejected in one of the sequential tests.[25] This is an interesting result since the CES speci- fication is gradually replacing the Cobb- Douglas specification in many areas of con- ceptual application of production function theory. Of course, there is an inconsistency in assuming labor aggregation on the one hand to test for capital aggregation, and re- jecting it on the other. Our primary pur- pose in this paper is to investigate the nature of the approximate tests for separa- bility. An obvious avenue for further re- search is to use the apparatus developed in this paper to test for homothetic separa- bility within the context of simultaneously disaggregated labor and capital inputs.

APPENDIX

PROOF of Proposition 1:

Berndt and Christensen (1973a) have shown that the Leontief conditions for weak separability of the translog function are

[25]Within the family of exact tests, the CES structure is not a testable hypothesis if one uses the translog function as the maintained hypothesis. Thus, Berndt and Christensen could not have ended up with this specification.

$$\alpha_i \gamma_{jk} - \alpha_j \gamma_{ik} + \sum_{m=1}^{N} (\gamma_{im}\gamma_{jk} - \gamma_{jm}\gamma_{ik}) \ln X_m = 0$$

where i, j index factors in the separable group to be aggregated and k indexes factors excluded from the separable group. A sufficient condition for this equation to hold is that $\gamma_{jk} = \gamma_{ik} = 0$. These are Berndt and Christensen's linear separability constraints. For $\gamma_{jk}, \gamma_{ik} \neq 0$ necessary and sufficient conditions for the above equation to hold for *every* input vector $x = (X_1, \ldots X_N)$, that is, *exactly*, are $\alpha_i\gamma_{jk} - \alpha_j\gamma_{ik} = 0$ and $\gamma_{im}\gamma_{jk} - \gamma_{jm}\gamma_{ik} = 0, m = 1, \ldots N$. This second set of conditions is Berndt and Christensen's non-linear separability restrictions and can be rewritten as $\alpha_i/\alpha_j = \gamma_{ik}/\gamma_{jk} = \gamma_{im}/\gamma_{jm}, m = 1, \ldots N$.

Now consider the 3-input case used in the text. The separability restrictions are $\gamma_{13} = \gamma_{23} = 0$ (linear restrictions) and $\alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23} = \gamma_{11}/\gamma_{21} = \gamma_{12}/\gamma_{22} = \theta$ (non-linear restrictions).

Substituting the linear separability conditions into the three-factor version of the translog production function (1), we obtain

$$\begin{aligned} \ln y &= \ln\alpha_0 + [\alpha_1 \ln X_1 + \alpha_2 \ln X_2] + \alpha_3 \ln X_3 \\ &\quad + \frac{1}{2}\left[\sum\sum_{i,j=1,2} \gamma_{ij}\ln X_i \ln X_j\right] + \gamma_{33}(\ln X_3)^2 \\ &= \ln\alpha_0 + \theta_G\left[\sum_i \beta_i \ln X_i + \frac{1}{2}\sum_i\sum_j \cdot \beta_{ij}\ln X_i \ln X_j\right] + \theta_H[\beta_3 \ln X_3 + \beta_{33}(\ln X_3)^2] \\ &\qquad i,j = 1,2 \end{aligned}$$

where $\alpha_i = \theta_G \cdot \beta_i$, $\alpha_k = \theta_H \cdot \beta_k$, $\gamma_{ij} = \theta_G\beta_{ij}$, $\gamma_{kk} = \theta_H\beta_{kk}$, $i,j = 1,2$; $k = 3$ and θ_G, θ_H are arbitrary constants.

Define $\ln G = \sum_i \beta_i \ln X_i + 1/2 \sum_i \sum_j \beta_{ij} \ln X_i \ln X_j$, $i,j = 1,2$, a translog aggregate input (up to an arbitrary scaling factor β_{OG}). Also define $H = \beta_3 \ln X_3 + \beta_{33}(\ln X_3)^2$, a translog aggregate (also up to an arbitrary scaling factor β_{OH}). Then $\ln y = \ln\alpha_0 + \theta_G \ln G + \theta_H \ln H$.

Now we consider the non-linear separability restrictions. Substituting these restrictions into equation (1) and utilizing the symmetry conditions $\gamma_{ij} = \gamma_{ji}$ we obtain

$$\begin{aligned} \ln y &= \ln\alpha_0 + \alpha_1\left[\ln X_1 + \frac{1}{\theta}\ln X_2\right] + \alpha_3 \ln X_3 \\ &\quad + \frac{1}{2}\gamma_{11}(\ln X_1)^2 + (\gamma_{11}/\theta)\ln X_1 \ln X_2 \\ &\quad + \frac{1}{2}\gamma_{22}(\ln X_2)^2 + \gamma_{13}\ln X_1 \ln X_3 \\ &\quad + (\gamma_{13}/\theta)\ln X_2 \ln X_3 + \frac{1}{2}\gamma_{33}(\ln X_3)^2 \\ &= \ln\alpha_0 + \alpha_1\left[\ln X_1 + \frac{1}{\theta}\ln X_2\right] + \alpha_3 \ln X_3 \\ &\quad + \frac{1}{2}\gamma_{11}\left[\ln X_1 + \frac{1}{\theta}\ln X_2\right]^2 \\ &\quad + \gamma_{13}\left[\ln X_1 + \frac{1}{\theta}\ln X_2\right]\ln X_3 \\ &\quad + \frac{1}{2}\gamma_{33}(\ln X_3)^2 \end{aligned}$$

since $\gamma_{11}(1/\theta)^2 = \gamma_{11}[\gamma_{22}/\gamma_{12}]^2 = \gamma_{12}^2/\gamma$ $\gamma_{22}^2/\gamma_{12}^2 = \gamma_{22}$.

Defining $\ln G = \theta_1 \ln X_1 + \theta_2 \ln X_2$ a $\ln H = \theta_3 \ln X_3$ (Cobb-Douglas aggrega we can write the separable production fu tion as

$$\ln y = \ln\alpha_0 + \beta_G \ln G + \beta_H \ln H + \frac{1}{2}\sum\sum_{i,j=G,H} \beta_{ij}\ln G \ln H$$

(a translog functi

where $\theta = \theta_1/\theta_2$, $\beta_G = \alpha_1/\theta_1$, $\beta_H = \alpha_3/$ $\beta_{GG} = \gamma_{11}/\theta_1^2$, $\beta_{GH} = \gamma_{13}/\theta_1\theta_3$, β_{HH} γ_{33}/θ_3^2.

PROOF of Proposition 4:

The quadratic approximation to (9) panded around $x^* = 1$ is

$$\hat{Q}_s = \beta_0 + \sum_{i=1}^{3} \beta_i \ln X_i + \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} \beta_{ij}\ln X_i \ln X_j$$

where $\beta_0 = f(x^*)$;

$$\beta_i = \frac{\partial f}{\partial G}\cdot\frac{\partial G}{\partial ln X_i}\bigg|_{x^*} \qquad i = 1,2$$

$$\beta_3 = \frac{\partial f}{\partial ln X_3}\bigg|_{x^*}$$

$$= \left[\frac{\partial f}{\partial G}\cdot\frac{\partial^2 G}{\partial ln X_i \partial ln X_j} + \left(\frac{\partial G}{\partial ln X_i}\right)\cdot\left(\frac{\partial G}{\partial ln X_j}\right)\cdot\frac{\partial^2 f}{\partial G^2}\right]_{x^*} \qquad i = 1,2$$

$$= \frac{\partial^2 f}{\partial ln X_3^2}\bigg|_{x^*}$$

$$= \beta_{3i} = \frac{\partial G}{\partial ln X_i}\cdot\frac{\partial^2 f}{\partial G \partial ln X_3}\bigg|_{x^*} \qquad i = 1,2$$

t can be seen that the above parameters sfy the constraint

$$_1/\beta_2 = \left|\frac{\partial G}{\partial ln X_1}\Big/\frac{\partial G}{\partial ln X_2}\right|_{x^*} = \beta_{13}/\beta_{23}$$

'he translog function (6) subject to the istraint $\alpha_1/\alpha_2 = \gamma_{13}/\gamma_{23}$ is identical to Since $\hat{Q}_s$ is a second-order approxima- n in the case of weak separability of uts 1 and 2 from 3; application of Defini- n 3 suffices to establish Proposition 4.

OOF of Proposition 4A:

Assuming linear homogeneity of f we can ply Euler's theorem to (10) to obtain

$$\frac{\partial ln y}{\partial ln X_i} \equiv 1 \quad \text{or}$$

$$\frac{\partial f}{\partial G}\sum_{i=1}^{2}\frac{\partial G}{\partial ln X_i} + \frac{\partial f}{\partial ln X_3} \equiv 1$$

ice,

$$\sum_{i=1}^{2}\frac{\partial G}{\partial ln X_i} \equiv \left(1 - \frac{\partial f}{\partial ln X_3}\right)\Big/\frac{\partial f}{\partial G}$$

te that the left-hand side is a function of X_1 and $ln X_2$ only, while the right-hand e is a function of G and $ln X_3$. Since the nt-hand side is independent of $ln X_3$ if it qual (identically) to the left-hand side, it s to be a function of G alone.

us
$$\sum_{i=1}^{2}\frac{\partial G}{\partial ln X_i} = \sum_{i=1}^{2}\frac{\partial G}{\partial X_i}\cdot X_i = h(G)$$

This is a necessary and sufficient condition that G be homothetic with regard to X_1 and X_2.

REFERENCES

E. R. Berndt and L. R. Christensen, (1973a) "The Translog Function and the Substitution of Equipment, Structures and Labor in U.S. Manufacturing 1929–68," *J. Econometrics*, Mar. 1973, *1*, 81–113.

——— **and** ———, (1973b) "The Internal Structure of Functional Relationships: Separability, Substitution, and Aggregation," *Rev. Econ. Stud.*, July 1973, *40*, 403–10.

——— **and** ———, "Testing for the Existence of a Consistent Aggregate Index of Labor Input," *Amer. Econ. Rev.*, June 1974, *44*, 391–404.

——— **and D. Wood,** "Technology, Prices, and the Derived Demand for Energy," *Rev. Econ. Statist.*, Aug. 1975, *57*, 259–68.

C. Blackorby, D. Primont, and R. Russell, "On Testing Separability Restrictions with Flexible Functional Forms," *J. Econometrics*, Mar. 1977, *5*, 195–209.

L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental Logarithmic Production Frontiers," *Rev. Econ. Statist.*, Feb. 1973, *55*, 28–45.

———, ———, **and** ———, "Transcendental Logarithmic Utility Functions," *Amer. Econ. Rev.*, June 1975, *65*, 367–83.

M. Denny and M. Fuss, "The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates," Inst. Policy Anal., working pap. 7506, Univ. Toronto 1975.

W. E. Diewert, "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function," *J. Polit. Econ.*, May/June 1971, *79*, 481–507.

M. A. Fuss, "The Demand for Energy in Canadian Manufacturing: An Example of the Estimation of Production Structures with Many Inputs," *J. Econometrics*, Jan. 1977, *5*, 89–116.

D. W. Jorgenson and L. J. Lau. "The Structure

of Consumer Preferences," *Annals Econ. Soc. Measur.*, Ap. 1975, *6*, 49–101.

L. Lau, "Applications of Duality Theory: A Comment," in David Kendrick and Michael Intriligator, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam 1974, 176–99.

Edmund Malinvaud, *Statistical Methods Econometrics*, Amsterdam 1966.

H. Uzawa, "Production Functions with C stant Elasticities of Substitution," *Econ. Stud.*, Oct. 1962, *29*, 291–99.

The Short-Run Dynamics of Prices and the Balance of Payments

By MARIO I. BLEJER*

The purpose of this study is to analyze theoretically and empirically the short-run behavior of prices and the balance of payments in a small open economy with a fixed exchange rate. Following the presentation of the theoretical framework, the experience of Mexico during 1950-73 is analyzed.

The model is based on what is known as the monetary approach to the balance of payments. This approach, as is frequently presented,[1] deals with long-run equilibrium situations when all relative prices are fixed. This long-run equilibrium view was applied to test empirically the theory for several countries,[2] using the assumption that there are no serious barriers to the international movement of capital and that all goods are traded.[3] These assumptions imply that goods and capital markets are perfectly arbitraged and, therefore, that price levels and interest rates in all the countries always move together.

However, as was pointed out by Alexander Swoboda, if a country has a low degree of capital mobility and a very large proportion of nontraded to traded goods, its speed of adjustment to monetary disturbances will be reduced, and even though the long-run convergence of inflation and interest rates will not be prevented,[4] the length of the disequilibrium period will be increased considerably. In order to analyze the experience of countries with those characteristics, a framework that allows for short-run intercountry variations in inflation rates is required.

In Section I a model with these properties is developed. This is done by using the distinction between traded and nontraded goods. The relative price of the two kinds of goods can be affected by domestic monetary policy, because the price of traded goods is exogenously determined for each country while the price of nontraded goods is subject to changes as a consequence of domestic policies.

As an application of the monetary approach to the balance of payments, the emphasis is on the interaction between the supply and the demand for money in explaining a country's balance of payments. The basic difference from the long-run formulation is that in this model monetary disequilibria affect not only the balance of payments but also, in the short run, the internal level of prices.

I. The Model

The two endogenous variables to be explained are the rate of domestic inflation, and the changes in the balance of payments as reflected in the balance of the money

*The Hebrew University of Jerusalem and Boston University This paper is based on my Ph.D. dissertation I wish to acknowledge my gratitude to R. Dornbusch, J A. Frenkel, A.C. Harberger, H.G. Johnson, and L A Sjaastad for their guidance and helpful suggestions I am also indebted to A.C. Porzecanski, to the managing editor of this *Review*, and to an anonymous referee for their valuable comments. The responsibility for remaining errors is solely mine.

[1]The classical references are the article by Harry Johnson and the works of Robert Mundell. For a general discussion of this approach see the volume by Jacob Frenkel and Johnson.

[2]A survey of the empirical evidence on the monetary approach to the balance of payments is presented by Stephen Magee The second part of the Frenkel and Johnson book collects several empirical papers on the topic

[3]An alternative assumption used in this context is that if some goods are intrinsically nontraded, by allowing a sufficiently long period of time the factors used in their production will be reallocated such that their earnings will continue to be comparable with those of the factors working in the traded-goods sector This is how price movements in both sectors are linked together.

[4]Long-run interest rate differentials will still be present if there exists imperfect substitution between foreign and domestic credit.

account (compensatory movement of reserves). Both variables are a function of what we call the *ex ante* excess flow supply of money. We assume full employment and, further, that monetary disturbances do not affect the level or the rate of growth of real income.[5]

The model is composed of the equilibrium conditions in the markets for money and for both types of goods, traded and nontraded. The formal analysis is presented in terms of discrete percentage changes of the variables over time. The letter D stands for the first-difference operator and the symbol * for the variables' percentage rate of change.[6]

The equations of the money sector are the following:

$$(1) \quad m_d = f(y, i)$$

$$(2) \quad M_d^* = P^* + m_d^*$$

$$(3) \quad M_s^* = a^* + \frac{Dc}{H} + \frac{Dr}{H}$$

where m_d the real demand for money is a function of real income y, and of the alternative cost of holding money i. The latter is the real interest rate plus the expected depreciation imposed by changes in the price level. M_d^* is the percentage rate of change of the demand for nominal cash balances and M_s is the nominal money supply; a is the money multiplier; c is the domestic credit outstanding, r the level of foreign-exchange reserves; and H is the monetary base ($H = r + c$);[7] P^* is the domestic rate of inflation, assumed to be a geometrically weighted average of the rate of change in the price of both goods:

$$(4) \quad P^* = \beta P_T^* + (1 - \beta) P_{NT}^*$$

where P_T is the price of traded goods,[8] P_{NT} is the price of nontraded goods, and β is the share of traded goods in total expenditure.

One of the postulates of the monetary approach to the balance of payments is that in a small open economy with a fixed exchange rate, the nominal money supply is beyond the control of the monetary authority. All that it can do is to determine the *ex ante* quantity of money by changing the domestic component of the base or manipulate variables under its control in order to change the value of the money multiplier. These actions, in conjunction with the flow demand for money that is generated by adjustments in the desired stock as a consequence of changes in real variables and expectations, create the *ex ante* excess flow supply of money to which the public reacts. It does so by changing the foreign component of the monetary base (through the balance of payments) and, in this short-run model, the rate of domestic inflation. The public, therefore, determines the *ex post* nominal quantity of money.

By subtracting the flow demand for nominal balances from the supply variables under the control of the monetary authority (a and c), we obtain g, a measure of the gap (in percentage terms) between the *ex ante* change in the money supply and changes in demand:

$$(5) \quad g = Dc/H + a^* - (P^* + m_d^*)$$

With regard to the goods' market, since an excess supply of money implies an excess demand for goods (both traded and nontraded), relative prices are sensitive to shocks from the monetary sector. This is so because an excess demand for nontraded goods results in an increase in their price while, in the small open economy we are considering, the price of traded goods is unaffected by changes in domestic demand.

[5]A model for an open economy in which real output fluctuates in the short run responding to monetary imbalance is presented by the author and Roque Fernandez.

[6]For any variable x, $Dx = (x_t - x_{t-1})$; $x^* = Dx/x_{t-1}$; and $Dx^* = x_t^* - x_{t-1}^*$.

[7]Equation (3) is derived as follows: since the monetary base H is identically equal to the domestic credit outstanding c and the foreign-exchange reserves of the central bank r, we have; $H = c + r$, and in rate-of-change: $H^* = DH/H = \xi(Dc/c) + (1 - \xi) Dr/r$, where $\xi = c/H$. Because $M_s = aH$, we obtain: $M_s = a^* + H^* = a^* + Dc/H + Dr/H$.

[8]Traded goods include exportables and importables whose prices are determined in world markets and thus exogenously for a small country. We assume therefore that the terms of trade are given.

If we further assume that an excess demand for nontraded goods varies monotonically with an excess demand throughout the economy, we can postulate a functional relationship of the following form between the relative prices and g, the *ex ante* monetary gap:

$$P_{NT}/P_T = ne^{\lambda g} \tag{6}$$

where n is a constant and λ stands for the elasticity of relative prices to monetary imbalance; λ takes values between 0 and ∞ and is mainly a function of the elasticities of substitution between traded and nontraded goods in production and in consumption as well as on the income elasticity of nontraded goods.

Equation (6) implies that for each level of the gap between the *ex ante* rate of money creation and the changes in the demand for money there exists a unique relative price of nontraded in terms of traded goods.[9] Differentiating equation (6) logarithmically we obtain:

$$P^*_{NT} - P^*_T = \lambda(Dg) \tag{7}$$

which implies that in a small open economy the rate of domestic inflation relative to the world rate is determined in the short run by the domestic excess flow supply of money.

Although it is possible to postulate a mechanism of lagged adjustments in the money market, we assume here for simplicity the existence of permanent *ex post* stock equilibrium, i.e., that in any period, the nominal quantity of money is equalized, *ex post*, with the desired quantity of nominal balances ($M_s = M_d$). This assumption requires the following flow equilibrium:[10]

$$M^*_s \equiv Dc/H + Dr/H + a^* = P^* + m^*_d \tag{8}$$

The system formed by (7) and (8) in conjunction with definitions (4) and (5) can be solved for the rate of domestic inflation and for the balance of payments (as reflected in the rate of change of international reserves). Both endogenous variables are expressed as functions of the world rate of inflation (assumed to be equal to the rate of change in the price of traded goods), the rate of change of the *ex ante* excess flow supply of money, and the past period's rate of inflation (reflecting lagged values of the former exogenous variables):

$$P^*_t = \frac{1}{1+\lambda(1-\beta)}(P^*_T)_t + \frac{\lambda(1-\beta)}{1+\lambda(1-\beta)}\left(D\left(\frac{Dc}{H} + a^* - m^*_d\right)\right)_t + \frac{\lambda(1-\beta)}{1+\lambda(1-\beta)}P^*_{t-1} \tag{9}$$

$$\left(\frac{Dr}{H}\right)_t = \frac{1}{1+\lambda(1-\beta)}(P^*_T)_t + \left(m^*_d - \frac{Dc}{H} - a^*\right)_t + \frac{\lambda(1-\beta)}{1+\lambda(1-\beta)}\left(D\left(\frac{Dc}{H} + a^* - m^*_d\right)\right)_t + \frac{\lambda(1-\beta)}{1+\lambda(1-\beta)}P^*_{t-1} \tag{10}$$

Equations (9) and (10) indicate that the distribution of the impact of monetary disequilibrium between the balance of payments and the domestic rate of inflation depends in the short run on the values of λ

[9] A formal derivation of an equation like (6) requires the specification of a utility function and a budget constraint, in addition to the supply conditions. Utility maximization gives us a set of demand equations that, in conjunction with the supply conditions, determine a complete model. However, in this framework and in order to discuss the dynamic properties of the model, some simplifying assumptions about the adjustment process are necessary. In the literature it is common to consider the Marshallian adjustment, where output adjusts to excess demand, or alternatively, the Walrasian adjustment where prices adjust to excess demand. Here, we postulate an adjustment process similar to the Walrasian adjustment but, instead of excess demand in the goods' market, we use excess supply in the money market.

[10] One of the criticisms of the monetary approach to the balance of payments is that its central formulations are based on transformations of an identity (the balance sheet of the central bank), and thus it has no empirical content. Nevertheless, it is clear from equation (8) that the identity refers only to the composition of the money supply, while the behavior of the balance of payments is derived from a set of postulates about functional forms in the monetary sector and an assumption about how the money market clears.

and β. If $\lambda = 0$, implying that relative prices are not affected by an excess supply of money, or if $\beta = 1$, which means that all goods are traded, we observe from (9) and (10) that:

$$(11) \qquad P^* = P_T^*$$

$$(12) \quad Dr/H = P_T^* + m_d^* - Dc/H - a^*$$

which is the conventional long-run model which assumes that the rate of domestic inflation is pegged to the world rate and that monetary disturbances affect only the balance of payments.

If, on the contrary, we set $\lambda = \infty$, the other extreme case is obtained: all the impact of monetary disequilibrium is on the domestic level of prices and the balance of payments is not directly affected.

The short-run effects of monetary disequilibrium will fall more heavily on the domestic price level and less on the balance of payments the higher is λ and the lower is the share of traded goods, β. The opposite holds for the effect of world inflation on the domestic rate of price increase. A policy implication of this observation is that governments can reduce the short-run impact on the balance of payments of a given monetary policy by imposing heavy restrictions to international transactions of goods and securities: prohibitive tariffs or quantitative restrictions will turn traded into nontraded goods, lowering the value of β and reducing the impact of a monetary disequilibrium on the balance of payments during the transition period. But these measures can only delay the full impact of the monetary changes on the balance of payments until the adjustment process is completed, and that will only be accomplished at the cost of a higher rate of domestic inflation during this adjustment period.

We turn now to the characteristics of equilibrium in this model. Equations (9) and (10) indicate that, if the world rate of inflation is unchanged, in order to maintain balance-of-payments equilibrium the monetary authority must keep the *ex ante* excess flow supply of money constant (i.e., $D(Dc/H + a^* - m_d^*) = 0$), increasing the domestic credit component of the monetary base permanently at a rate that exceeds the increase in the demand for money due to real factors (income growth, for example) by exactly the world rate of inflation. Assuming no changes in the money multiplier, the maintenance of equilibrium requires:

$$(13) \qquad \gamma_0 = \left(\frac{Dc}{H} - m_d^*\right) = P_T^*$$

where γ stands for the rate of *ex ante* excess flow supply of money. Since the coefficients of $(P_T^*)_t$ and P_{t-1}^* in equations (9) and (10) add up to unity we have:

$$(14) \quad P^* = P_T^* \qquad \text{and} \qquad Dr/H = 0$$

When no changes take place in the real sector to affect m_d^* and the rate of world inflation remains constant, an increase in the rate of domestic-credit creation raises the rate of the *ex ante* excess supply of money from γ_0 to γ_1, and generates a process of adjustment like the one depicted in Figure 1. The acceleration in the rate of domestic-credit expansion implies a positive value for $D(Dc/H - m_d^*)$ which raises the rate of change of nontraded goods' prices and, therefore, the domestic rate of inflation above the world rate. If the rate of domestic-credit creation is kept at this new level, the domestic rate of inflation will decelerate and converge to the world rate.

Since we do not allow here for stock disequilibrium, the *ex ante* excess flow supply of money is eliminated each period by a combination of price rises and reserves depletion. When the domestic rate of inflation converges to the world rate, the whole increase in the rate of domestic-credit creation that generated the disequilibrium is eliminated through the balance of payments.[11]

II. Empirical Implications: The Case of Mexico (1950–73)

The model presented in Section I is tested here using annual data for Mexico for the

[11] During the period of adjustment P_{NT} can exceed at some point, the value of γ_1 depending on whether $\lambda\beta > 1$ or $\lambda\beta < 1$. However, P^* will always be lower than γ_1 if λ is finite due to reserve depletion.

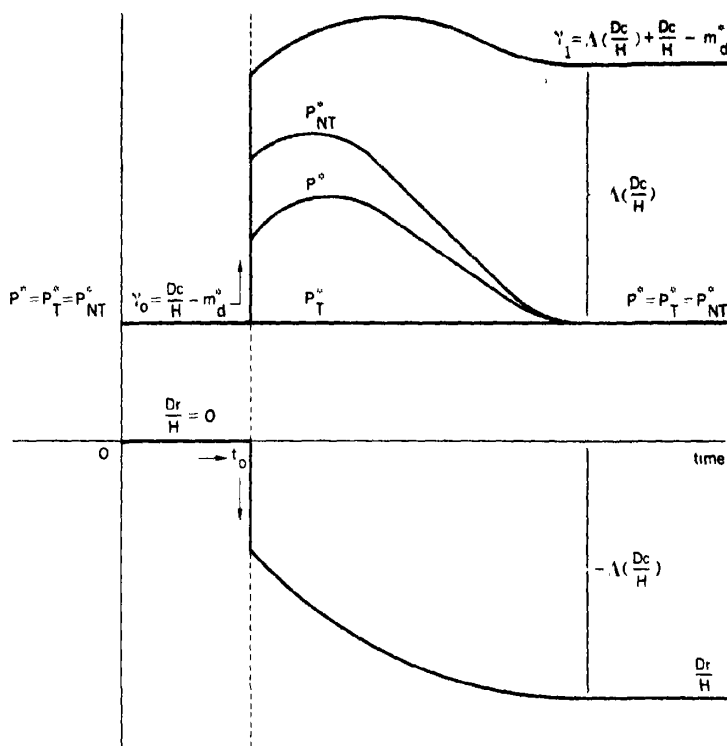



FIGURE 1. THE EFFECTS OF AN ACCELERATION IN THE RATE OF EXPANSION OF THE DOMESTIC-CREDIT COMPONENT OF THE MONETARY BASE

The curve γ represents the rate of *ex ante* excess flow supply of money ($Dc/H - m_d$). Since at the left of t_0, γ equals the world rate of inflation (P^*_T), the domestic rate of inflation (P^*) equals the world rate, and the balance of payments is in equilibrium ($Dr/H = 0$). At time t_0 the rate of domestic-credit expansion is increased by $\Delta(Dc/H)$ and maintained at the higher rate. The new path of γ (i.e., γ_1) first overshoots its new equilibrium rate since the rate of change of the real demand for money falls when the expectations of inflation increase. When these expectations fully adjust, γ converges to its new equilibrium rate. The figure shows that P^*_{NT} and P^* are first higher than P^*_T and later converge as the whole increase in domestic-credit expansion is eliminated through the balance of payments ($Dr/H = -\Delta(Dc/H)$).

period 1950–73.[12] Mexico is an interesting case study for the monetary approach to the balance of payments because of the long stability of the Mexican peso,[13] the freedom enjoyed by Mexicans to make international monetary transfers and payments, as well as the complete convertibility of the peso into foreign currency.[14] Moreover, because Mexican commercial policy was relatively unchanged during the period, exogenous disturbances in the relevant variables were not introduced.[15]

The main objective of this empirical section is to analyze the dynamics of the do-

[12]For a more complete empirical implementation of the model, see my dissertation.

[13]During the period analyzed here, the Mexican peso was devalued only once, in April 1954.

[14]Until 1968 the peso was fully convertible into gold but, following the creation of the two-tier gold market, the gold trade of the central bank with the public became somewhat restricted.

[15]The international movement of goods is restricted in several ways. However, what is important in testing the model is that the scheme of protection was kept relatively constant over the years.

mestic rate of inflation and of the rate of change of foreign-exchange reserves by estimating equations (9) and (10) of the theoretical model. However, to make the model more tractable to empirical estimation and to avoid econometric problems like the existence of a lagged endogenous variable in the right side of the equations, several transformations were made.

By using L, the lag operator,[16] and after some manipulations, equations (9) and (10) can be rewritten as:

$$(9') \quad P_t^* = \frac{1}{1+\lambda(1-\beta)}\left(\sum_{i=0}^{\infty} \alpha^i L^i (P_T^*)_t\right) + \frac{\lambda(1-\beta)}{1+\lambda(1-\beta)}\left(\left(\sum_{i=0}^{\infty} \alpha^i L^i - \sum_{i=0}^{\infty} \alpha^i L^{i+1}\right)\left(\frac{Dc}{H} + a^* - m_d^*\right)_t\right)$$

$$(10') \quad \left(\frac{Dr}{H}\right)_t = \frac{1}{1+\lambda(1-\beta)}\left(\sum_{i=0}^{\infty} \alpha^i L^i \left(P_T^* + m_d^* - a^* - \frac{Dc}{H}\right)_t\right)$$

where

$$\alpha = \frac{\lambda(1-\beta)}{1+\lambda(1-\beta)}$$

Expressions (9') and (10') indicate that the current rate of domestic inflation and the current rate of change of foreign-exchange reserves may be estimated as functions of polynomials of current and lagged rates of world inflation and of current and lagged rates of *ex ante* excess flow supply of money.

The rate of domestic inflation is a weighted average of those polynomials and the weights are functions of the elasticity of relative prices with respect to the excess flow supply of money λ, and of the share of traded goods in expenditure β. The lower is β and the higher is λ, the higher we expect the coefficient of the excess supply of money to be relative to the coefficient of P_T^*.

In the empirical estimations, the price of traded goods is approximated by the U.S. wholesale price index. The rate of change of the *ex ante* excess supply of money is calculated by adding the rate of change of domestic credit creation[17] to the rate of change of the money multiplier, and subtracting from this total an estimate of the rate of change of the real demand for money. To calculate these changes in money demand, a function like that given by equation (1) was estimated for three alternative definitions of money: M_1 (cash plus demand deposits); M_2 (M_1 plus time deposits); and M_3 (M_2 plus demand and time deposits denominated in foreign currency, which in Mexico are highly liquid and fully convertible, and therefore can be included in the definition of money). Two alternative measures of the opportunity cost of holding money were tested: the expected rate of inflation,[18] and a proxy for the nominal interest rate.[19]

The value of the polynomials is also a function of the values of λ and β. For given values of β, a smaller λ implies a lower weight for the past-year variable. Because the weights of the polynomial decline geometrically, lagged variables further in the past will have a lower influence on the current rate of inflation. The polynomials were constructed attaching alternative values to λ (for a given value of β, 0.60, calculated from the data). That produced weights for

[16]For any variable x_t, $L(x_t) = x_{t-1}$ and $L_i(x_t) = x_{t-i}$.

[17]Domestic credit is defined here as claims of the central bank on the federal government, on the publicly owned financial institutions, and on the private banking sector. Robert L. Bennet, Dwight S. Brothers and Leopoldo Solis, and Leslie W. Small studied in detail the characteristics of the Mexican financial sector.

[18]The proxy for expected inflation π^e was estimated by an error-learning process. The general form of the variable is:

$$\pi_t^e = \theta P_t^* + (1-\theta)\pi_{t-1}^e$$

Several alternative weights (θ) ranging from 0.2 to 0.9 were applied to the lag structure and the one reported here corresponds to $\theta = 0.6$, which maximized the R^2.

[19]The only series on interest rate data available for Mexico relates to a relatively small portion of the financial transactions carried out in the country. Only the commercial loans of private banks from their unrestricted asset portfolio are granted at rates that are approximate to the market-determined interest rates. These yields are the proxy used here.

TABLE 1—ESTIMATES OF EQUATION (9') FOR MEXICO, 1950–73: INFLATION MEASURED BY THE RATE OF CHANGE OF THE CONSUMER PRICE INDEX

Money Definition	Constant	$P^*_{U.S.}$ (1)	$\frac{DX_i}{X_i}$ (2)	$\frac{DX_e}{X_e}$ (3)	D_0 (4)	D_1 (5)	R^2	D.W.
M_1	−0.1009[a]	0.617	0.224		0.449	0.064[a]	0.690	1.88
	(−0.061)	(2.57)	(3.61)		(5.78)	(0.95)		
M_1	0.181[a]	0.533		0.290	0.447	0.068[a]	0.744	2.11
	(0.11)	(2.34)		(4.43)	(6.49)	(1.11)		
M_2	−0.747[a]	0.719	0.239		0.467	0.096[a]	0.670	1.78
	(−0.44)	(2.99)	(3.33)		(5.56)	(1.36)		
M_2	−0.448[a]	0.668		0.325	0.480	0.114[a]	0.720	2.03
	(−0.29)	(3.02)		(4.01)	(6.24)	(1.74)		
M_3	−0.577[a]	0 706	0.213		0.470)	0.095 [a]	0.693	1.79
	(−0.33)	(2.87)	(3.69)		(5.87)	(1.39)		
M_3	−0.182[a]	0.639		2.89	0.482	0.110	0.755	1.99
	(−0.11)	(2.84)		(4.64)	(6.82)	(1.80)		

Notes: Numbers in parentheses are *t*-values.
[a]Not significant at the 5 percent level.
(1) Distributed-lag polynomial of the *U.S* rate of inflation
(2) Distributed-lag polynomial of the *ex ante* excess flow supply of money using the interest rate in the estimations of the money demand
(3) Distributed-lag polynomial of the *ex ante* excess flow supply of money using the expected rate of inflation in the estimations of the money demand
(4) Dummy for 1954
(5) Dummy for 1955.

the first lagged year ranging from 0.1 to 0.9.[20] A number of trials were then performed to find the combination of polynomials that maximized the coefficient of correlation. For the inflation rate measured by the consumer price index, the R^2 was maximized with a weight of 0.6 for the first lag in the *U.S.* rate of inflation and a weight of 0.2 for the first lag in the *ex ante* excess flow supply of money. For the wholesale price index, the maximizing weights were 0.8 and 0.3, respectively. These results support the belief that the total adjustment to changes in domestic monetary conditions is completed faster than the adjustment to changes in the international price level.

The complete results for the inflation rate are presented in Table 1 (for the consumer price index) and Table 2 (for the wholesale price index). In addition to the polynomials, two dummy variables are introduced to account for the effects of the 1954 devaluation: one accounting for the effects of the devaluation on the rate of inflation during that year (D_0) and the other for its effects during the following year (D_1).[21]

All the estimated coefficients are highly significant except those of D_1, which seems to indicate that the major effect of the devaluation on the level of prices was exhausted after one year. Mexico is a country

[20]The steady-state gains of the polynomials as implied by the model are: $1 + \lambda(1 - \beta)$ for the rate of U.S inflation and 0 for the *ex ante* excess supply of money. The number of terms included was truncated when the weight corresponding to the following term was smaller than 0.01. The polynomials were then constructed and constrained to the implied gain by appropriately weighting the included terms in order to achieve the required sum.

[21]In order to consider the effects of devaluation on the domestic rate of inflation, equation (8) may be rewritten as follows:

$$(8') \qquad P^* = \beta(\omega^* + P^*_T) + (1 - \beta)P^*_{NT}$$

where ω^* is the rate of devaluation and P^*_T is the rate of change of the price of traded goods measured in foreign currency. Since the proxy for P^*_T used in our estimations is the rate of U S. inflation, the effect of devaluation on the price of traded goods is not accounted for. In order to do so, a variable measuring variations in the exchange rate must be included. In our particular case we use a dummy variable for ω^* because the Mexican peso was devalued only once during the whole period under analysis.

TABLE 2—ESTIMATES OF EQUATION (9′) FOR MEXICO, 1950–73: INFLATION MEASURED BY THE RATE OF CHANGE OF THE WHOLESALE PRICE INDEX

Money Definition	Constant	P^*_{US} (1)	$\frac{DX_i}{X_i}$ (2)	$\frac{DX_e}{X_e}$ (3)	D_0 (4)	D_1 (5)	R^2	D.W
M_1	−0.629[a]	0.375	0.225		0.368	0.083[a]	0.653	1.93
	(−0.47)	(2.80)	(3.61)		(5.10)	(1.27)		
M_1	−0.489[a]	0.352		0.258	0.366	0.090[a]	0.655	2.09
	(−0.387)	(2.77)		(3.55)	(5.16)	(1.38)		
M_2	−1.046[a]	0.435	0.201		0.363	0.109[a]	0.564	1.84
	(−0.06)	(2.95)	(2.58)		(4.33)	(1.49)		
M_2	−0.810[a]	0.423		0.228	0.368	0.119[a]	0.548	2.02
	(−0.55)	(2.95)		(2.36)	(4.25)	(1.56)		
M_3	−0.773[a]	0.414	0.182		0.360	0.102[a]	0.588	1.92
	(−0.53)	(2.91)	(2.83)		(4.53)	(1.43)		
M_3	−0.506[a]	0.397		0.213	0.365	0.111[a]	0.580	1.98
	(−0.373)	(2.90)		(2.70)	(4.54)	(1.53)		

Notes: Numbers in parentheses are t-values Columns (1)–(5): See Table 1.
[a]Not significant at the 5 percent level.

with almost unimpeded international mobility of capital and with relatively few restrictions to the movement of goods. In addition, its proximity to the United States increases the array of goods and services that are likely to be traded. Thus, we expect that the value of λ will be low and the value of β high, implying that in our model the weight for the P^*_{US} polynomial terms is higher relative to that of the *ex ante* excess flow supply of money. That is indeed confirmed by the estimations, since the coefficients of P^*_{US} are always considerably higher than those of the excess money supply.

In order to analyze the response of the balance of payments to monetary disequilibria, equation (10′) is estimated. As in the case of equation (9′), alternative values of λ are used to generate the weights for the polynomial. The value of R^2 is maximized when a weight of 0.35 is given to the first lagged year, which implies that a stock disequilibrium in the money market will create a flow process of adjustment in the balance of payments that will have a significant influence during a period of about three years.

The complete results for the rate of change of the foreign assets of the Bank of Mexico are presented in Table 3. The results indicate that the model has a good explanatory power in terms of R^2 and of the significance of the coefficients. In general the use of wider definitions of money gives a better fit than using M_1, as is expected from the nature of the model: if we are using the M_1 definition of money, a shift from demand deposits to another kind of deposits will be considered as a reduction in the demand for money but this shift, by itself, should not affect the flow of reserves.

As in the case of the rate of inflation, the devaluation appears to have had a significant effect on the rate of change of foreign-exchange reserves during the same year, but that effect is not statistically significant after a period of one year has elapsed.

III. Summary and Conclusions

In Section I a model is developed in order to analyze the impact of external influences and of domestic monetary disequilibria on the rate of inflation and on the balance of payments of a small fixed exchange rate economy. The model departs from other presentations of the monetary approach to the balance of payments since its main preoccupation is the study of the short-run characteristics of the adjustment process during which rates of inflation are allowed to differ among countries.

The central implication of the theoretical inquiry is that increasing the domestic-credit component of the monetary base at a

TABLE 3—ESTIMATES OF EQUATION (10′) FOR MEXICO, 1950–73: THE RATE OF CHANGE OF THE FOREIGN ASSETS HELD BY THE BANK OF MEXICO

Money Definition	Constant	$\left(P^*_{U.S.} - \frac{DX_i}{X_i}\right)$ (1)	$\left(P^*_{U.S} - \frac{DX_e}{X_e}\right)$ (2)	D_0 (3)	D_1 (4)	R^2	D.W.
M_1	2.363	0.455		0.500	0.156	0.751	1.90
	(2.18)	(4.10)		(4.29)	(1.41)		
M_1	2.155		0.481	0.513	0.146[a]	0.707	1.85
	(1.84)		(3.37)	(4.03)	(1.20)		
M_2	2.610	0.542		0.458	0.086[a]	0.808	2.04
	(2.72)	(5.23)		(4.40)	(0.85)		
M_2	1.656[a]		0.622	0.446	0.051[a]	0.772	2.09
	(1.14)		(4.47)	(2.85)	(0.45)		
M_3	2.38	0.507		0.468	0.096[a]	0.839	1.92
	(2.76)	(6.02)		(4.99)	(1.05)		
M_3	1.41		0.591	0.456	0.061[a]	0.811	1.88
	(1.54)		(5.29)	(4.41)	(0.60)		

Notes: The dependent variable is the rate of change of the foreign assets held by the Bank of Mexico, weighted by the share of the foreign component in the monetary base. Numbers in parentheses are t-values.

[a]Not significant at the 5 percent level.

(1) Distributed-lag polynomial of the difference between the rate of *U.S.* inflation and the rate of *ex ante* excess supply of money, using the interest rate in the estimations of the money demand.

(2) Distributed-lag polynomial of the difference between the rate of *U.S.* inflation and the rate of *ex ante* excess supply of money, using the expected rate of inflation in the estimations of the money demand.

(3) Dummy for 1954.

(4) Dummy for 1955.

rate too fast relative to the growth of the demand for money will result in a balance-of-payments deficit as well as in a short-run rate of inflation higher than that of the rest of the world. However, if the discrepancy between the supply of and the demand for money is maintained at this constant rate, the process of adjustment in the markets for goods, capital, and money will be completed, the domestic inflation rate will converge to that of the rest of the world, and the complete excess supply of money created *ex ante* will be eliminated through the balance of payments.

The empirical evaluation of the Mexican experience between 1950 and 1973 tends to support the predictions of the model. The Mexican rate of inflation, as measured by changes in both the consumer price index and the wholesale price index, is significantly explained by external inflation and by internal monetary disequilibria. In both cases the results indicate that the adjustment of prices to internal monetary imbalance is completed in a shorter period than the adjustment to external influences. The country's balance-of-payments performance is also well explained by the independent monetary variable. The lag structure indicates that a stock disequilibrium in the money market will create a flow process of adjustment in the balance of payments that will have relevant effects during a period of about three years. The 1954 devaluation of the Mexican currency had an effect on the rate of inflation and on the inflow of reserves during the same year, but these effects became insignificant after a period of one year.

REFERENCES

Robert L. Bennet, *The Financial Sector and Economic Development: The Mexican Case,* Baltimore 1965.

M. I. Blejer, "Money, Prices and the Balance of Payments. The Case of Mexico 1950–1973," unpublished doctoral dissertation, Univ. Chicago 1975.

——— **and R. B. Fernandez,** "On the Trade-off between Output, Inflation and the

Balance of Payments," Int. Trade Workshop, Univ. Chicago 1975.

Dwight S. Brothers and Leopoldo Solis, *Mexican Financial Development*, Austin 1966.

Jacob A. Frenkel and Harry G. Johnson, *The Monetary Approach to the Balance of Payments*, London 1975.

H. G. Johnson, "The Monetary Approach to the Balance of Payments Theory," in his *Further Essays in Monetary Economics*, London 1972, 229–49.

Robert A. Mundell, *International Economics*, New York 1968.

S. P. Magee, "The Empirical Evidence on the Monetary Approach to the Balance of Payments and Exchange Rates," *Amer. Econ. Rev.*, May 1976, *66*, 163–70.

L. W. Small, "An Analysis of Mexican Monetary Management, 1954–1969," unpublished doctoral dissertation, Univ. Indiana 1973.

A. K. Swoboda, "Monetary Policy Under Fixed Exchange Rates: Effectiveness, the Speed of Adjustment and Proper Use," *Economica*, May 1973, *40*, 136–54.

Measuring the Expected Real Rate of Interest: An Exploration of Macroeconomic Alternatives

By J. W. ELLIOTT*

At least since the time of Irving Fisher, it has been clear that nominal rates of interest differ from real rates not because of current or past price level changes but because of expected future price level changes. Accordingly, while nominal rates of return on fully discounted notes are observable magnitudes, expected real returns on the same notes are nonobservable. Nevertheless, real rates of return and real rates of interest are important concepts in the development of contemporary macroeconomic theory, particularly so as that theory develops richer theoretical roles for real rate and inflationary expectations measures. Furthermore, questions such as whether real rates are constant over time or are subject to systematic fluctuation through time reflect attributes of financial behavior that have important implications for understanding macroeconomic adjustment mechanisms. Even so, little attention has been given to the problem of forming and evaluating empirical measures of temporal movements in real rates of interest.[1]

The work done by Fisher and several contemporary followers has proceeded on the premise the expected real rates of interest are constant over time. This theory has usually been examined empirically by estimating a model of the form:[2]

$$(1) \qquad r_t = \rho + \pi_t$$

$$(2) \quad \pi_t = \sum_{i=0}^{n} W_i(p_{t-i} - p_{t-i-1})_t + \mu_t$$

where r_t is the nominal one-period rate of interest as of t, ρ is the expected real rate assumed to be invariant over time, π_t is the current expected rate of inflation over the term of r_t, p_t, is the *log* of the price level at t, W_i are distributed lag coefficients, and μ_t is an error term. Empirical estimates of (2) have produced measures of W_i that imply extremely long lags, a finding that leads to considerable suspicion about its usefulness.[3] A severe theoretical restriction on such analysis has recently been pointed out by Thomas Sargent (1973b), who shows essentially that a peculiar set of restrictions upon the parameters in an *IS-LM* framework are required for the results of estimating π_t by estimating the W_i in (2) to be statistically reliable.[4] For these reasons, and since this paper seeks to measure and examine temporal properties of the real rate, the Fisherian framework of expres-

*Professor of economics and business administration, University of Wisconsin-Milwaukee I wish to thank the Urban Research Center and the Northwestern Mutual Life Insurance Company for financial support in this project.

[1] Insightful discussions of some of the conceptual problems involved in macroeconomic models involving price expectations have been given by Thomas Sargent (1973a, b), and by Sargent and Neil Wallace. An interesting discussion of most of the problems associated with forming rational expectations estimates of price level changes has been given recently by John Rutledge However the only attempt to define and evaluate a measure of the time-series movement in inflationary expectations of which I am aware is a study by Patric Hendershott and James Van Horne. Their study is greatly limited in its usefulness by the assumption that risk premiums between bond and equity markets are constant over time.

[2] Recent examples of this can be found in papers by William Gibson (1970) and William Yohe and Denis Karnosky. Fisher also ran similar tests.

[3] Papers by Phillip Cagan, Marc Nerlove, and Sargent (1973a, b) raise such doubts and offer possible reasons for the findings obtained in some of the previous work

[4] Sargent's objection appears to be associated with the general property of most theoretical macroeconomic constructs that either (a) both π_t and ρ are endogenous variables, or (b) π_t is exogenous while ρ is endogenous. In either case, attempts to determine π_t that assume ρ constant over time are likely to produce misleading results upon empirical analysis of most real world data in which this restriction is not likely to be present.

sions (1) and (2) is employed only as a baseline against which to evaluate other less problematical approaches.[5]

Other work by Stephen Turnovsky and to a certain extent by David Pyle and William Gibson (1972) has explored the extent to which the survey data of J. A. Livingston, a syndicated columnist for the *Philadelphia Bulletin*, are useful to represent inflationary expectations, from which the real rate derives. Their evidence has not suggested the survey approach is superior to statistical means of measuring inflationary suggestions.[6] In this paper, the empirical problem of measuring real rates of interest is addressed by statistical means under the premise that information can be gleaned from market behavior that is at least as useful as information that could be provided by survey participants. In addition, the real rate is assumed to be a time-subscripted variable that influences and is influenced by the operation of macroeconomic markets. Accordingly, we write

$$(1') \qquad r_t = \rho_t + \pi_t$$

and initially seek to explain ρ_t by appealing to macroeconomic theory. Alternative models for the real rate are defined based on neo-Keynesian and loanable funds macroeconomic frameworks. Estimates of the real rate are formulated from Muthian premises for several alternative models. The performance of each resulting measure of the expected real rate as an estimator of the future realized real rate is used to judge its efficiency as a representation of rational real rate expectations. The results show a basically neo-Keynesian model specification for the equilibrium real rate of interest coupled with a monetary explanation of expected inflation rates to be the most efficient vehicle for the measurement of expected real rates of interest. A measure of the expected real rate is calculated from this model structure over the period 1960–74, and is found to be an unbiased estimate of the *ex post* real rate, subject to substantial short-run shifts over time, and largely independent of fluctuations in real output in its time-series behavior.

[5]Eugene Fama has recently tested the Fisherian assumption that the real rate is constant and reported positive evidence in this regard.

[6]A reason often expressed for the lack of explanatory dominance of survey data is that the feelings of respondents that are expressed in survey forms do not coincide with the influences that motivate their actions. In this regard, George De Menil and Surjit Bhalla have recently concluded essentially the same thing in finding that a measure of inflationary expectations developed from the *Michigan Survey of Consumer Finances* explain wage changes about as well or perhaps somewhat better than a group of indirect measures previously proposed.

I. Measuring the Expected Real Rate

The real rate is an expectation widely held by investors. John Muth's concept of such an expectation is that it is rational if it is equal to the prediction of relevant theory. To apply this attractive notion to measuring the expected real rate requires (a) specification of what constitutes "relevant theory" vis-à-vis real rate fluctuations, and (b) extracting predictions of the real rate from the theoretical model.

The question of what constitutes relevant theory for such an expectations measure can be usefully examined in terms of the efficiency of alternative theoretical constructs in using available information. A particular expectations model is more efficient than another if it more completely reflects the information contained in current macroeconomic variables. Accordingly, an expectations measure that reflects current information more fully than another measure is "rational" in that it produces more accurate predictions than other alternatives. Thus, the efficiency of an expectations measure in the use of information can be judged by the accuracy of predictions drawn from it compared with other candidate measures.

In constructing such an expectations measure for the real rate, several plausible theoretical alternatives are present. First, a class of macroeconomic constructs is possible in which real income, prices, and real rates are assumed to be endogenous variables jointly determined by commodity, money, and labor market adjustment processes. We refer to these as neo-Keynesian

A second class of macroeconomic constructs from which a rational expectations real rate model can be formed assumes real income to be exogenous to the market clearing mechanism in commodity, money, and bond markets. Principle endogenous variables in these models are prices and real interest rates. We refer to these models as neoclassical-loanable funds.

Both neo-Keynesian and loanable funds constructs involve real and nominal rates of interest; thus both require a theoretical way to account for fluctuations in the inflationary expectation that connects real and nominal rates. At least two theoretical alternatives are present in this regard: (a) a monetary explanation of inflationary expectations, and (b) a factor market explanation that relies upon wage rate dynamics. In addition, both constructs involve long and short rates and thus implicitly involve both long and short inflationary expectations. In what follows, we assume both long and short inflationary expectations to be approximately equal and thus capable of being characterized by a single measure.

In this paper, specific econometric representations of neo-Keynesian and loanable funds models are developed, as are monetary and labor market models for inflationary expectations. When each of the models of inflationary expectations are coupled with each of the macroeconomic models, the result is four plausible alternative models of the expected real rate. The efficiency of these alternatives in the sense of Muthian rationality is judged by their relative performance in real rate predictions.

II. A Neo-Keynesian Framework

To provide a neo-Keynesian explanation of ρ_t, a simple model is employed that is nearly identical to that recently proposed by Sargent (1973a). It is given as follows:

Aggregate Supply

$$(3)\quad y_t - d_t = a_1(p_t - {}_tp^*_{t-1}) + v_{1,t} \qquad a_1 > 0$$

Aggregate Demand

$$(4)\quad y_t - d_t = b_1\rho_t + b_2z_t + v_{2,t} \qquad b_1 < 0,\ b_2 > 0$$

Monetary Balances

$$(5)\quad m_t - p_t = c_1y_t + c_2r_t + v_{3,t} \qquad c_1 > 0,\ c_2 < 0$$

where y_t, p_t, m_t, and d_t are the natural logarithms of real national income, the price level, the nominal money supply, and the level of capacity real income, respectively; where ${}_tp^*_{t-1}$ is the level of prices expected to prevail at time t as of time t − 1; and where z is a group of exogenous or predetermined determinants of aggregate demand. Specific definitions of the statistical series used to measure these and all other variables are given in Appendix A.

In this model, the aggregate supply schedule relates the gap of actual real output from full-employment output to unanticipated price level changes. This reflects the model of aggregate supply studied intensively by Robert Lucas (1973a, b), and by Lucas and Leonard Rapping in which producers consistently perceive their product price increases as not being accompanied by factor price increases and thus are willing to expand output proportionately.

The aggregate demand schedule given by expression (4) is a simple *IS* curve in which real interest rates and a vector of exogenous demand determinants (z) are assumed to determine the gap between aggregate demand and full-employment supply. The exogenous demand determinants are (a) government expenditure and net exports, both assumed to be autonomous demand components, and (b) a distributed lag on past output change, assumed to be a determinant of aggregate investment demand.

Monetary equilibrium is given by expression (5) as a condition where demand for real money balances, driven by real income and nominal interest rates, is equated with the supply of real balances.[7]

In the model given by expressions (3)–

[7]This expression differs from Sargent's (1973) model by not assuming that $C_1 = 1$. In a later paper, Sargent

(5), y_t, p_t, r_t, and ρ_t are endogenous while m_t, d_t, and z_t are exogenous. To complete the model specification, it is necessary to provide an additional equation in the endogenous variables.[8] For our purpose, it is appropriate to specify an additional expression for the inflationary expectation, $\pi_t = r_t - \rho_t$. With a structural expression for π_t added, the model provides a complete explanation for the four endogenous variables. A model for the real rate results by solving expression (4) for ρ_t as follows:

(6)
$$\rho_t = b_1^{-1}(y_t - d_t) - b_2 b_1^{-1} z_t - b_1^{-1} v_{2,t}$$

and by replacing y_t by its first stage reduced-form estimate, $\hat{y}_t$, to avoid the simultaneity bias that otherwise would accompany the estimation of coefficients in (6) by single equation methods. The specific formulation of z we employ is:

$$(7) \qquad z_t = GX_t + \sum_{i=1}^{12} W_i \Delta y_{t-i}$$

where GX is the level of government real spending plus real net exports, and where the W_i are distributed lag weights assumed to lie along a third-order polynomial. This formulation of z assumes that the level of government outlays and an expectional type investment accelerator are principal influences on the position of the IS curve. The value of $\hat{y}_t$ is obtained by regressing y_t upon the highest causal order predetermined variables in the model structure which consists of all unlagged exogenous variables. Thus:

$$(8) \qquad \hat{y}_t = k_0 + k_1 \cdot d_t + k_2 \cdot m_t + k_3 \cdot z_t + k_4 \cdot J_{\pi,t}$$

where $J_{\pi,t}$ is the vector of exogenous determinants of π_t according to the model for inflationary expectations we add to complete the neo-Keynesian model specification. To specify $J_{\pi,t}$, we initially consider a monetary explanation of inflationary expectations.

also uses the form of this equation given by expression (5). See Sargent and Wallace.

[8]This model can be considered complete in its present form if inflationary expectations are taken to be exogenous.

III. A Monetary Model of Inflationary Expectations

The monetary model developed here for explaining price expectations has similarities to a monetary model for "consistent" expectations developed by Allen Walters in the course of an interesting discussion of the use of rational expectations to defend purely autoregressive models of inflationary expectations. Assuming the ratio of monetary velocity to real income to be approximately constant in the short run produces the conclusion that the equilibrium level of prices, P_t^e, is proportional to the nominal money supply, as:

$$(9) \qquad P_t^e = \alpha \cdot M_t$$

$$\text{or} \qquad \log P_t^e = \log \alpha + \log M_t$$

which we denote as:

$$(10) \qquad p_t^e = a + m_t$$

Actual price levels are seen as continually adjusting towards p_t^e according to a partial adjustment structure:

$$(11) \quad p_{t+1} - p_t = [1 - \beta][p_{t+1}^e - p]$$

which implies

$$(12) \quad p_{t+1} = [1 - \beta] \sum_{i=0}^{\infty} \beta^i p_{t+1-i}^e$$

We can generalize this to:

$$(13) \qquad p_{t+1} = \sum_{i=0}^{\infty} \mu_i p_{t+1-i}^e$$

where $\Sigma \mu_i = 1$.

Differencing both sides of (13) gives:

$$(14) \quad p_{t+1} - p_t = \sum_{i=0}^{\infty} \mu_i [p_{t+1-i}^e - p_{t-i}^e]$$

Under rational expectations and the monetarist rule given by (10), we may replace p^e in (14) by the expected value of m, denoted by $E(m)$, plus the constant value a and obtain the expected value of the price level, $E(p)$.

It follows that:

$$(15)\quad E[p_{t+1}] - p_t = \sum_{i=0}^{\infty} \mu_i [E(m_{t+1-i}) - m_{t-i}] + e_{1,t}$$

where $e_{1,t}$ is an error term. We assume that the expected value of the monetary growth term one quarter ahead is proportional to the most currently observed value, i.e., that $E(m_{t+1})$ is $(1 + \Phi)(m_t)$ where Φ is the expected normal growth rate in the *log* of nominal money supply. Since $E[p_{t+1}] - p_t = \pi_t$ according to our definition, we have:

$$(16)\quad \pi_t = \Phi \sum_{i=0}^{\infty} \mu_i \cdot m_{t-i} + e_{1,t}$$

as the monetary explanation of expected inflationary rates. In the empirical tests, we truncate this lag beyond 11 quarters. Inserting the right-hand side of expression (16) into expression (8) in place of $J_{\pi,t}$ and fixing the error at its mean of zero gives:

$$(17)\quad \hat{y}_t = k_0 + k_1 \cdot d_t + k_2 \cdot m_t + k_3 \cdot z_t + k_4 \phi \sum_{i=0}^{11} \mu_i \cdot m_{t-i}$$

as the first-stage equation for obtaining $\hat{y}_t$. Estimation of the second-stage structural equation (6) using $\hat{y}_t$ enables a rational expectation estimate of ρ_t to be calculated as the prediction of this model. We denote this estimate as the neo-Keynesian-monetary explanation of ρ_t. We now consider alternative rational expectations models.

IV. A Neoclassical-Loanable Funds Model

As an alternative to the neo-Keynesian macroeconomic framework, an essentially neoclassical loanable funds explanation of the real rate is employed. The structure used has some similarities to a model recently presented and analyzed by Panayotas Korliras, which views the interest rate to be the direct product of supply of and demand for real purchasing power (or real bonds) in financial markets.

The total supply of real bonds B^s consists of a business component BB^s that associates with the demand for real investment goods by business, and a government and foreign component GXB^S.

As long as debt securities comprise a steady proportion of total investment funds of firms, the business component of bond supply is proportional to an aggregate investment function. Taking this and the following relationships to be linear, primarily for convenience, we have:

$$(18)\quad BB_t^s = d_1 \rho_t + L_1 \cdot \Delta Y_{r,t} + W_{1,t} \qquad d_1 < 0;\ L_1 > 0$$

where $Y_{r,t}$ is the level (not *log*) of real aggregate income; L_1 is a lag operator representing a time-distributed accelerator effect on investment demand identical to that used in the neo-Keynesian structure, ρ_t is the expected real rate as before, and $W_{1,t}$ is an error term.

The Government Bond Supply GXB_t^s is equal to the net government borrowings from the public. From the standpoint of a Treasury/Federal Reserve sources and uses of funds statement, GXB_t^s is a residual that equates total sources and uses of funds, essentially as follows:

$$GXB_t^s = Def + Ex + \Delta Bal - \Delta M_b$$

where *Def* is the gross government deficit including interest and transfers, *Ex* is government purchases of foreign exchange, ΔBal is the change in Treasury Operating Balances, and ΔM_b is the change in the monetary base. The values of *Def* and ΔBal are usually assumed to be related to aggregate income. In the present loanable funds context, where aggregate income can be taken as largely independent of the interest rate, these terms are similarly independent. With ΔM_b and *Ex* also normally thought to be independent, GXB_t^s can be assumed to be exogenous in the loanable funds model.[9] Specifying the lag in (18) to be polynomial of third degree and 12 quarters in length gives

[9] In the empirical tests, the individual source-use components are not separated, but instead the net borrowings variable is taken as a single exogenous entity.

$$(19) \quad B^s = d_1\rho_t + \sum_{i=1}^{12} X_i \Delta Y_{r,t-i} + GXB_t^s + W_{1,t}$$

The total demand for real bonds B^d consists of private sector and public sector components. The private sector component includes demand for bonds by consumer and business sectors, and a component of demand related to the money-creating powers of commercial banks. Having earlier identified GXB_t^s as a net supply of bonds due to public sector effects, the total net demand for real bonds is just equal to the private sector demand.

Private sector real savings S_r is allocated in part to changes in holdings of real money balances and in part to a demand for bonds. The total savings flow S_r is related to aggregate income according to a simple neoclassical saving function:

$$(20) \quad S_{r,t} = e_0 + e_1 Y_{r,t-1} + e_2\rho_t + W_{2,t} \qquad 0 < e_1 < 1, e_2 > 0$$

in which the level of real income last (quarterly) period is taken to be an appropriate income magnitude for the loanable funds framework. This retains an important postulate of neoclassical loanable funds theory that current real income is determined independently of current interest rates and prices. Interest rate movements are primarily important in partitioning current real income into consumer and investment flows in the process of regulating the flow of money capital through financial markets. The use of one lagged income reflects this idea and also preserves the statistical integrity of our measures in the case where real income and interest rates are in fact jointly determined variables.

Taking the flow of current real saving as fully allocated to either changed levels of desired real money balances, ΔMH_t, or changed levels of consumer, business, and bank security holdings, PB_t^d, gives the following:

$$(21) \quad PB_t^d = S_{r,t} - \Delta MH_t$$

Combining (21) and (20) gives:

$$(22) \quad PB_t^d = e_0 + e_1 Y_{r,t-1} + e_2\rho_t - \Delta MH_t + W_{2,t}$$

In addition, it is assumed that banks initiate portfolio adjustments by alterations in the inside money component that change the flow of security demand for a given level of consumer and business saving. Since this is done by changes in the level of efficiency at which reserves are employed within the constraints imposed by the Federal Reserve, a component of bank-initiated demand for bonds is possible that is related to changes in the ratio of reserves to required reserves held by commercial banks as follows:

$$(23) \quad CBB_t^d = f_1 \Delta RD_t + W_{3,t} \quad f_1 > 0$$

where CBB_t^d is the demand for securities due to the portfolio adjustment of banks and where RD_t is the ratio of total reserves to required reserves. The total demand for bonds B^d is the sum of consumer, business, and bank-initiated demands, or adding (22) and (23):

$$(24) \quad D^d = e_0 + e_1 Y_{r,t-1} + e_2\rho_t + f_1 \Delta RD_t - \Delta MH_t + W_4$$

Changes in desired real money balances are assumed to be governed by a conventional demand for money rule in which changes in real income and in nominal rates of interest are principal influences. To this we add a component of demand representing the change in transactionary demand for real money balances by governments, giving:

$$(25) \quad \Delta MH_t = g_1 \Delta Y_{r,t-1} + g_2 \Delta r_t + g_3 \Delta G_t + W_{5,t} \qquad g_1 > 0; g_2 < 0; g_3 > 0$$

where ΔG_t is the change in real government outlays. In this expression, the most recently attained real income change is again used as in expression (20) for the same reasons as discussed previously.

To complete the loanable funds model, we assume the bond market is continually cleared in the sense that $B_t^d = B_t^s$ at all t and that changes in the level of desired real money balances are brought into equality

with changes in the supply of real money balances ΔMS_t at each t. In the latter case, we assume that changes in the nominal money supply have both an exogenous and an endogenous component while current price level is endogenous. As the ratio of the two, the real money supply can thus be taken to be endogenous in the present analysis. With $Y_{r,t-1}$ given, adjustments in r_t and in P_t occur until

$$\Delta MS_t = \Delta MH_t$$

while at the same time:[10]

$$B_t^s = B_t^d = B_t$$

With these conditions imposed, the loanable funds structure is complete once a specification for π_t is added. The endogenous variables are r_t, B_t, ΔMS_t, and $\pi_t = r_t - \rho_t$ while the structural equations are expression (19) for B^s, expression (24) for B^d, expression (25) for $\Delta MH_t = \Delta MS_t$, and the expression we add for π_t.

From this framework, we can construct an expression for the real rate of interest by setting expression (19) for B^s equal to expression (24) for B^d, and solving for the value of ρ_t, as follows:

$$(26)\quad \rho_t = K_0 + K_1[GXB_t + \Delta MS_t] + K_2 \cdot Y_{r,t-1} + K_3 \cdot \Delta RD_t + \sum_{i=1}^{12} X_i \Delta Y_{r,t-i} + W_{6,t}$$

where the coefficients $K_0, \ldots, K_3$ and X_i are combinations of the d, e, f, and g in which $K_1 > 0$, $K_2 < 0$, $K_3 > 0$, and $\Sigma X_i > 0$ due to the restrictions on the original coefficients. To estimate K_i, the endogenous variable ΔMS_t is replaced by its first-stage proxy $\Delta \hat{MS}_t$, which is obtained from the following reduced form expression:[11]

$$(27)\quad \Delta \hat{MS}_t = h_0 + h_1 GXB_t^s + h_2 \cdot Y_{r,t-1} + h_3 \cdot \Delta RD_t + h_4 \cdot \Delta G_t + h_5 \Delta Y_{r,t-1} + h_6[J_{\pi,t}]$$

where as before $J_{\pi,t}$ is a vector of exogenous variables associated with the expression we employ for π_t to complete the model structure.

Using the monetary model of inflationary expectations (expression (16)) to specify $J_{\pi,t}$ gives the completed first-stage expression for MS_t as:

$$(28)\quad \Delta \hat{MS}_t = h_0 + h_1 GXB_t^s + h_2 \cdot Y_{r,t-1} + h_3 \cdot \Delta RD_t + h_4 \cdot \Delta G_t + h_5 \cdot \Delta Y_{r,t-1} + h_6 \phi \sum_{i=0}^{11} \mu_i \cdot m_{t-i}$$

When the structural expression (26) is estimated in the second stage using $\Delta \hat{MS}$ in place of ΔMS, a rational expectation estimate of ρ_t can be obtained as the prediction of this model. We denote this estimate as the loanable funds-monetary explanation of the expected real rate.

V. An Alternative Inflationary Expectations Model

As an alternative to associating inflationary expectations with changes in the monetary aggregate, adjustments in labor market variables can be assumed to reveal the prevalent inflationary expectation in a different way. In a Walrasian-type labor market adjustment mechanism of the general type recently discussed by Jerome Stein, the rate of change in the money wage rate can be decomposed into a real and an inflationary component as follows:

$$(29)\quad \frac{\Delta W_{t+1}}{W_t} + \frac{\Delta (W/P)_{t+1}}{(W/P)_t} + \frac{\Delta P_{t+1}}{P_t}$$

where W is the money wage rate and P the price level. At time t, expectations about the values of W and P in t + 1 replace their actual values, giving:

$$(30)\quad \frac{E(\Delta W_{t+1})}{W_t} = \frac{E(\Delta (W/P)_{t+1})}{(W/P)_t} + \frac{E(\Delta P_{t+1})}{P_t}$$

Under rational expectations, the expecta-

[10] More complex disequilibrium models are overlooked at this point in favor of a tractable empirical structure.

[11] As in the neo-Keynesian case, lagged endogenous variables are excluded due to their low causal order along with the lag distribution on $\Delta Y_{t,t-i}$.

tional measures in expression (30) are governed by current relevant information and are equal to the one-period ahead predictions produced by this information. Since $E(\Delta P_{t+1})/P_t = \pi_t$ according to our earlier definition, a model of rational inflationary expectations can be formed from (30) by specifying determinants of expected money and real wage changes. In the *U.S.* economy of the 1960's and 1970's with which we are concerned, regular increases in money rates have been both expected and realized by wage earners, generally regardless of product or labor market conditions, i.e., money wages give the appearance of being largely exogenously determined over this period. Indeed, when a quadratic trend and a first-order autoregressive pattern are removed from the time path of the money wage rate over this period, the resulting deviation from trend shows no relationship to either labor market or product market conditions.[12] The character of the time path of money wages during the period of our analysis suggests the representation of expected money wage rate changes as a distributed lag on recent past actual money wage rate changes. Accordingly, we propose:

$$(31) \quad \frac{E[\Delta W_{t+1}]}{W_t} = \sum_{i=0}^{8} m_i(\log W_{t-i} - \log W_{t-i-1}) \qquad \Sigma m_i > 0$$

If money wage rates are largely autoregressive, the major burden of short-run real wage adjustments to changing equilibria falls upon product prices. The equilibrium real wage $(W/P)^e_t$ is that which clears labor markets. When the market real wage rises above $(W/P)^e_t$, the re-equilibrating process emphasizes adjustments in product prices, regardless of whether levels of production are at or below full employment. When the market real wage falls below $(W/P)^e_t$, demand for output exceeds full-employment production. If money wage rate changes continue to be independent of labor market conditions in this circumstance, then product shortages occur as producers are unable to produce at levels called for by their labor demand functions. As money wage rates adjust upward over time in their normal fashion, the demand for labor subsides toward full-employment levels.

This real wage adjustment process implies a time-distributed adjustment mechanism. Assuming real wage expectations are influenced in part by the expected growth rate in the equilibrium real wage and in part with a view towards the operation of a time distributed adjustment mechanism gives the following:

$$(32) \quad \frac{E[\Delta(W/P)_{t+1}]}{(W/P)_t} = E[\log(W/P)^e_{t+1}] - \log(W/P)^e_t + \sum_{i=0}^{8} n_i[\log(W/P)^e_{t-i} - \log(W/P)_{t-i-1}] \qquad \Sigma n_i > 0$$

In expression (32), the distributed lag term measures the adjustment process of the market real wage to the equilibrium levels according to our previous discussion. The $E[\log(W/P)^e_{t+1}] - \log(W/P)^e_t$ term is the expected change in the equilibrium real wage. In the United States, technologically based advances in labor productivity that regularly exceed increases in the supply of labor lead to a normally increasing equilibrium real wage. To obtain a time-series representation of $(W/P)^e$, data have been taken on the actual real wage, the workage population, and upon labor productivity for the 27 quarters over the 92-quarter period 1952.1 to 1974.4 in which the level of real national income was within 1 percent of its full-employment potential as measured by the U.S. Commerce Department.[13] These 27

[12] In a simple trend analysis of the money wage rate over 1960–74, a quadratic trend is found to improve the fit of either a linear or log-linear trend. Generalized least-squares estimation is necessary to reduce the strong autocorrelation in the residuals from this fit. The deviation from this trend component is found to have a coefficient of determination (R^2) of .04 relative to the gap of actual real *GNP* from potential *GNP*, and an R^2 of .01 with respect to the unemployment rate.

[13] The percent of potential has been rounded to the nearest whole percent. The full-employment quarters are 1952.1–1952.4, 1953.3, 1953.4, 1955.1–1956.2, 1965.3, 1965.4, 1967.1–1969.3, 1973.1, and 1973.2.

data points provide an indicaton of the movement in the equilibrium real wage over time. Using this full-employment sample, the real wage was regressed on the workage population and upon labor productivity. A good statistical fit resulted.[14] The resulting equation is used to interpolate and extrapolate to fill out the entire time-series and thus give a crude estimate of $(W/P)^e$. Estimated in this way, it appears safe to rule out feedback from current real interest rates to $(W/P)^e$. Accordingly, we take $(W/P)^e_t$ as exogenous for purposes of determining current inflationary expectations. To specify the determinants of $E[log(W/P)^e_{t+1}]$, we make the assumption that an average long-run rate of growth in the log of $W/P)^e_t$, r, from current levels is expected each time period, so that

(33)
$$E[log(W/P)^e_{t+1}] = (1 + r)(log(W/P)^e_t)$$

Combining expressions (30), (31), (32), and (33) gives the complete labor market model of inflationary expectations:

(34)
$$\pi_t = \sum_{i=0}^{8} m_i(log\ W_{t-i} - log\ W_{t-i-1}) + r[log(W/P)^e_t] - \sum_{i=1}^{8} n_i[log(W/P)^e_{t-i} - log(W/P)_{t-i-1}]$$

With respect to the first-stage equations in the neoclassical and loanable funds models, the right-hand side of expression (34) constitutes an alternative to the monetary model for the specification of $J_{\pi,t}$. When it is combined with the neo-Keynesian equations (6) and (8), a third rational expectations estimate of ρ_t results. We denote this combination as the neo-Keynesian-labor market model. When expression (34) is combined with the loanable funds equations (26) and (27), a fourth rational expectations estimate of ρ_t results. We denote this as the loanable funds-labor market model.

Finally, the Fisherian formulation of the real rate as the difference between the nominal rate and a distributed lag on past inflation rates (as given by expression (2)) provides a fifth way an estimate of ρ_t can be derived. Judging the predictive performance of the Fisherian model compared to the other models now given is of some importance since the Fisherian model has been used as an expectations generating mechanism in a number of different studies, including papers by Richard Roll, Gibson (1970), Leonall Anderson and Keith Carlson, Martin Feldstein and Otto Eckstein, Yohe and Karnosky, Turnovsky, and Franco Modigliani and Robert Shiller. Furthermore, a recent paper by Charles Nelson shows that a purely autoregressive expectations model will provide estimates that are efficient in the sense of producing minimum predictive error only for a particular and highly simplified underlying macroeconomic model structure.[15] Accordingly, the Fisherian formulation of ρ_t provides a useful baseline against which to compare the relative advantage of the structural models now proposed.

VI. Empirical Tests

The following alternative rational expectations models of the real rate have now been given:

a) neo-Keynesian ρ_t–labor market π_t
b) neo-Keynesian ρ_t–monetary π_t
c) loanable funds ρ_t–labor market π_t
d) loanable funds ρ_t–monetary π_t
e) Fisherian autoregressive model (equation (2))

[14] The equation is as follows:

$$(W/P)^e_t = \underset{(0.97)}{-16.32} + \underset{(2.81)}{.577}\,(Pop)_t + \underset{(2.71)}{.374}\,(out/MH)_t$$

$R^2 = .985$
$S.E. = 0.57$
$D\,W = 2.07$

$\rho_1 = 1.352$ $\quad n = 27$
$\rho_2 = -.354$ $\quad F(2,24) = 788.8$

where $(Pop)_t$ is the 16 and over populations and (out/MH) is output per man-hour in the private sector, ρ_1 and ρ_2 are autoregressive parameters in generalized least-squares estimating procedure employed due to high OLS autocorrelation.

[15] Where the true expectations generating mechanism is driven by several exogenous disturbances, Nelson shows that estimating expectations by these influences will always be more efficient than the autoregressive structure.

The 60-quarter period 1960–74 is chosen as the data base for the present analysis, largely on the notion that the economic relationships upon which real rates are based are likely to be more nearly constant over a period of this general span than over a substantially longer period.

A special problem in the 1960–74 data is the occurrence of the price controls of 1971–73, which may have had an effect on inflationary expectations over this period. This is handled by adding to each of the models a dummy variable with a value of zero during precontrol and postcontrol quarters and a value of one during the quarters in which controls were in force. In this way, the partial impact of controls can be ascertained and accounted for in the analysis.

The models are initially estimated over the 1960–74 period to enable in-sample predictive tests. In these tests, the dependent variable is the *ex post* one-quarter ahead real rate, ρx_t. The value of ρx_t is obtained by subtracting from a 90-day nominal composite spot rate at t the actual rate of price change from t to t + 1. This *ex post* value can be characterized as being composed of an expected portion earlier identified as ρ_t and an unexpected portion, ϵ_t, i.e., $\rho x_t = \rho_t + \epsilon_t$. Since ρ_t is equal to the calculated value from the associated economic model, the use of least squares estimation procedures constrains the in-sample estimates of ρ_t to be unbiased, i.e., $E[\rho x_t] = E[\rho_t]$; $E[\epsilon_t] = 0$. In addition, the ϵ_t are assumed to contain no predictable pattern, i.e., they are assumed to be white noise. Under these conditions, the principal information provided by the in-sample test is a comparison of the accuracy of each model in terms of the size of the residual variation each exhibits over the sample period. Table 1 shows these results. The neo-Keynesian model for the real interest rate completed by the monetary explanation of π_t provides the lowest residual standard error and the lowest average absolute error of the alternatives examined. The neo-Keynesian monetary model is accordingly the most efficient characterization of expectations among the alternatives examined according to the in-sample tests.[16] Further, each of the four macroeconomic models clearly dominates the autoregressive model in terms of residual variation, a result that suggests the information associated with the macroeconomic explanations of π_t is considerably more efficient than the information contained in past inflation rates.[17]

TABLE 1—ONE-PERIOD AHEAD PREDICTIONS OF ρ_t OBTAINED ON AN IN-SAMPLE BASIS[a] 1960–1974

	Root Mean Square Error	Average Absolute Error
1) Neo-Keynesian-Wages	1.16	0.95
2) Neo-Keynesian-Monetary	1.09	0.88
3) Loanable Funds-Wages	1.12	0.92
4) Loanable Funds-Monetary	1.12	0.92
5) Fisherian Autoregressive	4.10	3.93

[a]Shown in percent

[16]Two additional real rate models can be obtained by alternatively using the monetary and labor market models of π to obtain estimates of the rationally expected inflation rate. By making the assumption that the real rate is essentially a stationary stochastic process of a random walk nature, the rational expectations estimates of π can be subtracted from the nominal rate to obtain estimates of ρ_t. This approach essentially restricts all the arguments in the real rate structural equations in both macroeconomic models to have zero coefficients. Fitting the monetary model over the 1960–74 period gives an average absolute error of 1.06 percent, while the wages model produces an average absolute error of 1.18 percent. Both are higher than the worst of the macroeconomic explanations. An F-test of the significance of the restriction implied by the monetary model that the coefficients in the real rate structural equation are zero are 1.72 for the neo-Keynesian equation and 1.57 for the loanable funds equation. Both are significant at the 5 percent level. Comparable F-statistics for the wages model of 1.64 and 1.85 show the same result. Thus, the restrictions upon the macroeconomic real rate structure that are implied by the presumption that the real rate is a random walk are not confirmed in the data. The more sophisticated real rate model structure has significantly increased explanatory power.

[17]As an additional type of test relating to Nelson's work, the statistical significance of adding the autoregressive model structure to the best fitting neo-Keynesian monetary model was determined. It was found that the information contained in past inflationary rates adds nothing to the explanation of inflationary rates provided by the neo-Keynesian monetary model ($F = 1.17$ for the autoregressive distributed lag). How-

TABLE 2—CONDITIONAL POSTSAMPLE PREDICTIONS
(Shown in Percent)

	Prediction[a] Standard Error	Average Absolute Error	Mean Error (T-value)[b]
1) Neo-Keynesian-Wages	1.95	1.63	−0.84 (2.29)[c]
2) Neo-Keynesian-Monetary	1.69	1.44	−0.47 (1.47)
3) Loanable Funds-Wages	1.90	1.59	−0.40 (1.23)
4) Loanable Funds-Monetary	1.91	1.60	−0.37 (1.20)
5) Fisherian Autoregressive	3.37	2.43	−0.55 (0.87)
6) Composite	1.94	1.61	−0.53 (1.27)

[a]The residual in this analysis is defined as the predicted value minus the subsequent actual value.
[b]The t-value given is the t-statistic resulting from the null hypothesis that the mean residual is zero.
[c]Significant at the 5 percent level

A 28-quarter postsample prediction test of each model has been made to allow a comparison of the one-period ahead predictive powers of the various models. The models are initially estimated over the 20-quarter period 1960–67 and used for conditional predictions over the four quarters of 1968. They are then reestimated over 1960–68 for conditional predictions of 1969, and so forth with the last estimation interval being 1960–73 for a 1974 prediction.[18] The resulting predicted real interest rates measure expectations one quarter ahead. To evaluate predictive accuracy and prediction bias, they are compared with the subsequent *ex post* real rate. In addition, a composite prediction is calculated as the arithmetic average of other predictions to help examine the extent to which various individual predictions are offsetting. These results are summarized in Table 2, which gives the postsample prediction standard error, the average absolute error, and the mean prediction error. Each individual prediction is contained in Appendix B. For the mean prediction error, a t-statistic has been constructed for the null hypothesis that no prediction bias is present, i.e., that the population mean error is zero. The table shows that significant prediction bias at the 5 percent level is present in the case of the neo-Keynesian wages explanation of the real rate but is not demonstrable in the case of any of the other models tested. However, all models produced predictions having a mean lower than the actual value, largely due to a substantial underprediction of the 1974 inflationary rate. In fact, the composite of all predictions shown in the sixth line of the table shows the mean prediction to be low by about 1/2 of a percent over the sample period, an amount not significantly different from zero on our t-tests.

The lowest residual standard error and average absolute prediction error in the postsample prediction tests is produced by

ver, in the reverse case, it is also found that the structural explanation does not add significantly to the explanation of fluctuations in π_t provided by the autoregressive model ($F = 2.67$). These tests are similar to those proposed in a recent paper by Michael Hamburger and E.N. Platt as weak form and semistrong form tests of the efficiency of information use. My result seems to indicate that the two models are broadly similar in their explanatory capability. Given the Table 1 results, the macroeconomic model is the better of the two alternatives by virtue of its clearly lower standard error.

[18]The wage-price control dummy is excluded from the estimations until the 1960–71 estimation at which sufficient data under controls is present to warrant its inclusion. The 1971 predictions did not have a mechanism to account for controls, but were not apparently influenced to a great degree by the condition.

the neo-Keynesian monetary model, consistent with the in-sample tests. The three other macroeconomic approaches, 1), 3), and 4) in Table 2, rank next and about equally in residual variation. The autoregressive model is the least effective in postsample predictions by a considerable margin. These rankings are entirely consistent with those of the in-sample tests. Little value is seen to attach to a linear composite of the various predictions, as this predictor is among the least efficient of those considered. This result suggests most of the model prediction errors are in the same direction and are not importantly offsetting.

On the basis of Tables 1 and 2, the neo-Keynesian explanation of the real rate of interest completed by the monetary explanation of inflationary expectations stands out as both the best fitting model over the 1960–74 period and as the most efficient for conditional predictions over the 1968–74 period. It accordingly makes the most effective use of available theory and data. Accordingly, the real interest rate measured by this model is the best representation of Muthian rational expectations among the alternatives considered. At the other extreme, the autoregressive formulation makes the least effective use of current data and is the poorest representation of expectations we have considered. This comparatively poor showing raises questions about the continued use of the autoregressive model in empirical work, when more efficient alternatives are apparently available.

VII. Temporal Properties of the Expected Real Rate

My tests suggest that the neo-Keynesian-monetary explanation of the real rate leads to comparatively accurate and unbiased predictions. In the usual case where the coefficients in this model cannot be assumed to be stable over an extended time period, the measured real rate must be based on the postsample predictions of a frequently reestimated model.[19] This approach is followed in estimating the expected real rate of interest over the 1960-74 period. Postsample predictions over 1968–74 were previously obtained for the postsample tests reported upon in Table 2 by annual reestimates of the model. To these, similar postsample predictions are obtained in the same way and added to complete the 1960–67 portion of the expected real rate series.[20]

Table 3 shows some of the properties of the completed expected real rate series along with similar statistics for the *ex post* real rate. As seen, the mean expected real rate over the period is 1.44 percent compared to 1.23 percent for the *ex post* series. As the last column in Table 3 indicates, the difference is not statistically significant on a *t*-test. Thus, the expected real rate gives unbiased estimates of the *ex post* rate in postsample predictions over the 60 quarters 1960–74. The variance of the expected real rate series is also seen to be significantly less than the variance of the *ex post* series, as an *F*-ratio between the two is significant beyond 1 percent. This is an anticipated result, since expectations are usually thought to ignore the unsystematic "blips" that accompany *ex post* results.

That the expected real rate reaches a minimum value that is negative is a result of further interest. If the estimated real rate had been derived from a long-maturity nominal rate, a negative expected rate is somewhat unreasonable aside from the possibility it arose from measurement error. However, for the 90-day nominal rate used in this study, Baumol-Tobin type motives for portfolio positions in notes vis-à-vis cash take precedence over investment motives related to the supply of longer term investment capital. For the former motives the nominal rate of interest is the relevant

[19]This is the most general approach. If we assume the coefficients are entirely stable temporally, we can make all the predictions from a single set of coefficients that may have been estimated over a previous period or on an in-sample basis.

[20]To complete the 1960–67 portion of the series, the neo-Keynesian monetary model was estimated over 1952.1–1959.4 for past sample predictions of 1960.1–1960.4. Then, it was reestimated over 1952.1–1960.4 for the prediction of 1961.1–1961.4, and so forth.

TABLE 3—EXPECTED AND *Ex Post* REAL RATE
(Shown in Percent)

	Expected Real Rate (ρ_t)	*Ex Post* Real Rate (ρx_t)	Prediction Error: Expected *Ex Post* ($\rho_t - \rho x_t$)
Mean	1.44	1.23	0.21 (1.03)[a]
Minimum	−0.90	−3.30	−3.61
Maximum	3.49	4.24	4.41
Standard deviation	1.05	1.53	1.62
Average absolute error	—	—	1.28
F(*ex post* vs. expected variance) = 2.12[b]			

[a]Number in parentheses is the calculated t-statistic for the test of the null hypothesis that the mean prediction error is zero
[b]Significant beyond the 1 percent level.

input since the return on cash is the negative of the inflationary rate.

Accordingly, a nonzero level of equilibrium short-term securities holdings in lieu of money balances should be expected even at negative real rate expectations, so long as nominal short-term rates are nonzero. Since both of our macroeconomic models use the same short-term interest rate in both real and monetary sectors, they both assume an essentially invariant term structure relationship between short and long rates. Although we cannot measure this relationship in the present study, it is quite plausible that the measured spread between real short rates and real long rates is positive and sufficiently large to enable real expected long rates to be uniformly nonzero. The observance of negative extremes for the expected short rate thus can be consistent with a uniformly positive expected real long rate.

The Fisherian premise that the expected real rate is substantially constant over time is not supported by inspection of the estimate of this rate now obtained. The range of over 4 percentage points and standard deviation of greater than one percentage

TABLE 4—EXPECTED REAL RATE RELATIONSHIPS

	R^2	Regression F-Value	*GLS RHO*	*D W*
(A) $\rho_t = .0111 + .1802\dot{Y}_{r,t}$ (2.57) (2.49)	.10	6.21[a]	.85	1.77
(B) $\rho_t = .0131 - .0001 Y_{r,t}$ (0.51) (0.04)	.01	0.00	.87	1.70
(C) $\rho_t = .0199 - .1801\pi x_{t-1}$ (4.32) (3.42)	.17	11.69[b]	.85	1.65
(D) $\rho_l - \rho_t = .0240 - .5811\pi x_{t-1}$ (1.37) (4.42)	.25	19.55[b]	.92	1.99
(E) $\rho_l - \rho_t = -.0037 + .6154\dot{Y}_{r,t}$ (0.80) (3.61)	.18	13.07[b]	.66	2.19[b]

Note: All equations are estimated with generalized least squares (*GLS*) with a first-order autoregressive structure due to high autocorrelation. Numbers in parentheses are computed t-values. Symbols are as defined in the text.
[a]Significant at the 5 percent level for (1,58) d.f.
[b]Significant at the 1 percent level for (1,58) d.f.

point given in Table 3 indicate substantial temporal variability.

A further prima facie analysis of the expected real rate can be obtained by examining covariant patterns between ρ_t and measures of real output and current inflation. Results of this analysis are given in Table 4. In equations (A) and (B), ρ_t is regressed on both the rate of real output change and real output levels. A weak relationship with the rate of real output change and no relationship with output level is observed. From this, a dominant business cycle influence upon the expected real rate is not apparent in the estimated series. Expected real rates appear to be substantially independent of current real output levels.

In Table 4 (C), some relationship appears to exist between the temporal pattern of the real rate and the current actual rate of inflation, πx_{t-1}.[21] The negative and statistically significant nature of this relationship suggests that expected real rates are systematically lowered when the most current realized rate of inflation is increasing, presumably because the same real market and monetary factors impacting upon the current *ex post* inflationary rate are also impacting upon expectations about future inflationary rates and thus upon current expected real returns.

In Table 4(D) and (E), the prediction error in the real rate is regressed on both *ex post* inflation rates, and rates of change in real output. In both cases, a statistically significant covariance is found. In case of (D), increases in realized inflation rates tend to associate with smaller real rate prediction errors and vice versa. In the case of (E), the rate of real output change tends to directly associate with real rate prediction errors. Both these somewhat peculiar patterns could profitably be the subject of further future analysis.

[21] Note that the value of x in period t is the realized inflationary rate over the period t to t + 1. Thus x_{t-1} is the realized inflationary rate from t − 1 to t.

APPENDIX A

Data Measurement and Definitions (All Quarterly)

All symbols are as used in the text. Variables defined in the text by identity are not repeated here. Unobservable variables are also not included.

r_t = composite short-term rate, equal to the arithmetic average of the 90-day Treasury Bill rate, the rate on 90-day bankers acceptances, and the short-term commercial paper rate. Source: Federal Reserve.

p_t = *log* of the consumer price index. Source: U.S. Bureau of Labor Statistics (*BLS*).

y_t = *log* of the constant dollar *GNP*. Source: U.S. Office of Business Economics.

d_t = *log* of capacity real output level. Source: *NBER Economic Indicators.*

GX_t = *log* of the sum of total federal and state and local spending plus net exports, measured in constant dollars. Source: U.S. Office of Business Economics.

m_t = *log* of money supply − $M1$ definition. Source: Federal Reserve.

GXB_t = total constant dollar net borrowing from the public.

RD_t = ratio of total reserves to required reserves in the banking system. Source: Federal Reserve.

G_t = total government spending in constant dollars. Source: U.S. Office of Business Economics.

$Y_{r,t}$ = constant dollar *GNP*. Source: U.S. Office of Business Economics.

MS_t = money supply ($M1$) in constant dollars. Source: Federal Reserve.

W_t = *log* of average hourly wage in private sector. Source: *BLS*.

P_t = consumer price index. Source: *BLS*

ρx_t = *ex post* real interest rate, equal to r_t minus the rate of inflation from t to t + 1. Source: Federal Reserve for r_t and *BLS* for price levels.

APPENDIX B
Prediction Results

	Ex Post Real Rate	Predicted Rates (percent)				
		Neo-Keynesian		Loanable Funds		Fisherian
		Wages	Monetary	Wages	Monetary	Model
1968.1	1.46	0.88	0.94	0.44	0.53	1.03
.2	1.03	0.55	0.66	0.14	0.03	1.24
.3	0.54	0.49	0.62	−0.15	−0.14	0.71
.4	0.78	0.46	0.61	−0.03	−0.16	0.62
1969.1	0.10	0.71	0.80	0.72	0.52	0.95
.2	1.55	0.78	0.83	0.50	0.37	1.17
.3	2.05	0.80	0.78	0.56	0.40	1.54
.4	1.63	0.81	0.69	0.46	0.37	1.16
1970.1	2.19	1.26	0.80	0.92	0.84	0.62
2	3.27	1.12	0.66	0.56	0.57	−0.34
3	1.57	0.82	0.43	0.13	0.20	−1.27
4	2.37	0.93	0.39	−0.04	−0.03	−3.19
1971.1	0.22	1.47	1.53	1.42	1.37	−1.24
.2	0.90	1.50	1.60	1.48	1.46	−0.48
.3	3.02	1.40	1.56	1.27	1.26	0.47
.4	1.04	1.58	1.66	1.11	1.14	−0.16
1972.1	1.10	1.50	1.83	1.28	1.32	1.30
.2	0.74	1.49	2.01	1.71	1.74	2.18
.3	0.81	1.72	2.25	1.70	1.70	3.00
.4	−0.93	1.81	2.44	1.47	1.57	3.88
1973.1	−2.26	1.04	1.01	0.78	0.77	3.35
2	−1.51	1.34	1.06	1.34	1.21	4.56
3	−0.13	1.72	1.10	1.38	1.24	6.87
4	−3.30	2.02	1.12	1.04	0.95	6.06
1974.1	−2.79	1.46	−0.67	0.18	0.23	−1.06
2	−2.21	1.77	−0.45	0.91	1.00	−0.64
3	−1.37	2.10	−0.28	1.13	1.21	−0.76
4	0.75	2.48	−0.22	1.36	1.33	−3.42

REFERENCES

L. C. **Anderson and K. M. Carlson,** "A Monetarist Model for Economic Stabilization," *Fed. Reserve Bank St. Louis Rev.*, Apr. 1970, *52*, pp. 7–21.

Phillip Cagan, *Determinants and Effects of Changes in the Stock of Money, 1875–1960*, New York 1965.

G. **De Menil and S. S. Bhalla,** "Direct Measurement of Popular Price Expectations," *Amer. Econ. Rev.*, Mar. 1975, *65*, 169–80.

E. F. **Fama,** "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269–82.

M. S. **Feldstein and O. Eckstein,** "The Fundamental Determinants of the Interest Rate," *Rev. Econ. Statist.*, Nov. 1970, *52*, 363–76.

Irving Fisher, *The Theory of Interest*, New York 1930, 399–451.

W. E. Gibson, "Price-Expectations Effects on Interest Rates," *J. Finance*, Mar. 1970, *25*, 19–34.

———, "Interest Rates and Inflationary Expectations: New Evidence," *Amer. Econ. Rev.*, Dec. 1972, *62*, 854–65.

M. J. Hamburger and E. N. Platt, "The Expectations Hypothesis and the Efficiency of the Treasury Bill Market," *Rev. Econ. Statist.*, May 1975, *57*, 190–99.

P. H. Hendershott and J. C. Van Horne, "Ex-

pected Inflation Implied by Capital Market Rates," *J. Finance*, May 1973, *28*, 301–14.

P. G. Korliras, "A Disequilibrium Macroeconomic Model," *Quart. J. Econ.*, Feb. 1975, *89*, 56–80.

R. E. Lucas, (1973a) "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1973, *4*, 103–124.

———, (1973b) "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

——— **and L. A. Rapping,** "Real Wages, Employment, and Inflation," *J. Polit. Econ.*, Sept./Oct. 1969, *77*, 721–54.

F. Modigliani and R. J. Shiller, "Inflation, Rational Expectations, and the Term Structure of Interest Rates," *Economica*, Feb. 1973, *40*, 12–43.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

C. R. Nelson, "Rational Expectations and the Predictive Efficiency of Economic Models," *J. Bus., Univ. Chicago*, July 1975, *48*, 331–43.

M. Nerlove, "Distributed Lags and Unobserved Components in Economic Time Series," in William Fellner et al., eds, *Ten Economic Studies in the Tradition of Irving Fisher*, New York 1967.

D. H. Pyle, "Observed Price Expectations and Interest Rates," *Rev. Econ. Statist.*, Aug. 1972, *54*, 275–80.

R. Roll, "Interest Rates on Monetary Assets and Price Index Changes," *J. Finance Proc.*, May 1972, *27*, 250–78.

John Rutledge, *A Monetarist Model of Inflationary Expectations*, Lexington 1974.

T. J. Sargent, (1973a) "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment," *Brookings Papers*, Washington 1973, *2*, 429–72.

———, (1973b) "Interest Rates and Prices in the Long Run: A Study of the Gibson Paradox," *J. Money, Credit, Banking*, Feb. 1973, *5*, 385–449.

——— **and N. Wallace,** "'Rational' Expectations, the Optional Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

J. L. Stein, "Unemployment, Inflation, and Monetarism," *Amer. Econ. Rev.*, Dec. 1974, *64*, 867–87.

S. J. Turnovsky, "Empirical Evidence on the Formation of Price Expectations," *J. Amer. Statist. Assn.*, Dec. 1970, *65*, 1441–54.

A. A. Walters, "Consistent Expectations, Distributed Lags, and the Quantity Theory," *Econ. J.*, June 1971, *81*, 273–81.

W. P. Yohe and D. S. Karnosky, "Interest Rates and Price Level Changes, 1952–1969," *Fed. Res. Bank St Louis Rev.*, Dec. 1969, *51*, 18–38.

Board of Governors of the Federal Reserve System, *Fed. Res. Bull.*, various issues, Washington.

National Bureau of Economic Research (*NBER*), *Economic Indicators*, various issues, Washington.

U.S. Bureau of Labor Statistics (*BLS*), *Wages and Prices*, various issues, Washington

U.S. Office of Business Economics, *Surv. Curr Bus.*, various issues, Washington.

The Deterrent Effect of Capital Punishment: Another View

By PETER PASSELL AND JOHN B. TAYLOR*

In a recent paper in this *Review*, Isaac Ehrlich found a negative relationship between the use of the death penalty and homicide rates, over the three and a half decades from 1935–69 in the United States. This empirical finding, in apparent conflict with previous studies on the subject (see Thorsten Sellin, Robert Dann), takes on special meaning because (a) it is the first test of the capital punishment deterrence hypothesis to use econometric estimation techniques, and (b) its publication comes at a time when legislatures and the courts in both the United States and Canada are reshaping public policy on the use of the death penalty. Hence the importance of a second look at Ehrlich's results.

Such a reexamination, described below, suggests that the time-series model and data used by Ehrlich permit no inference about the deterrent effect of capital punishment on homicide. The data indicate that the parameters of Ehrlich's model are extremely sensitive to the choices of included explanatory variables and the functional form of the model; when recast in equally plausible alternative structures, the Ehrlich model fails to generate a deterrent effect for executions. Moreover, Ehrlich's methodological approach to the question, estimation of a single structural equation relating executions to murders, sharply limits the policy implications of the estimated coefficients.

Section I briefly describes Ehrlich's murder rate function and tests the model for parameter stability over the sample period. Section II considers his choice of explanatory variables and functional forms, and examines sensitivity of his estimate to these choices. Section III examines the pitfalls of inferring an execution-murder tradeoff from a single estimated equation.

I

A representative sample of the empirical results reported by Ehrlich in support of the deterrence effect of capital punishment is his equation (1), Table 4, which is displayed as our equation (1) in Table 1. The logarithm of the *U.S.* annual murder rate $(Q/N)^0$ is explained by the logarithm of three deterrence variables (the clearance rate for murder P_a^0, the conviction rate for murder $P_{c/a}^0$, and the ratio of current executions to one-year lagged murder convictions $(PXQ_1)_{-1}$), as well as by the logarithm of five other explanatory variables (the labor force participation rate L, the fraction of the population between 14 and 24 years of age A, an estimate of permanent real per capita income Y_p, the unemployment rate U, and an exponential time trend e^T). The equation is estimated by a two-stage procedure in which P_a^0 and $P_{c/a}^0$ are treated as endogenous variables, and are first regressed on current and lagged values of the predetermined variables, lagged values of all the endogenous variables, and a group of otherwise excluded exogenous variables (real police expenditure per capita $XPOL$, real government expenditure per capita $XGOV$, and the fraction of nonwhites in the population NW). In the second stage an iterative procedure is used to estimate the first-order serial correlation coefficient along with the coefficients of the endogenous and predetermined variables.

Ehrlich presents several versions of this same model (Tables 3 and 4, p. 410), among which the major differences are the use of alternatives to $(PXQ_1)_{-1}$ as empirical sur-

*Columbia University. An early version of this paper was written before the Ehrlich appendix appeared; it uses an alternative data set, available on request from the authors. This paper was listed as a Columbia Workshop discussion paper; it was published in a modified and less technical version in 1976. The conclusion and findings are similar despite the different data set. We would like to thank Phoebus J. Dhrymes and David Kennett for their comments and assistance.

TABLE 1

Equation	Constant	P_a^0	$P_{c/a}^0$	$(PXQ_1)_{-1}$	L	A	Y_p	U	T	P_{63}	C_{63}
(1) 1935–69	−4 060	−1 247	−0.345	−0.066	−1.314	0.450	1 318	0 068	−0.046		
SSR = .048	(−1.00)	(−1.56)	(−3 07)	(−3 33)	(−1 49)	(2.20)	(4 81)	(2 60)	(−6 54)		
$\hat{\rho}$ = .059											
(2) 1935–69	−5 536	−0 973	−0 375	−0 062	−1 620	0 565	1 363	0 065	−0 046		
SSR = 049	(−1 40)	(−1 26)	(−3 10)	(−3 01)	(−1 87)	(2.33)	(4 52)	(2.34)	(−5 86)		
$\hat{\rho}$ = 204											
(3) 1935 62	−7 219	0 124	0 236	−0 008	−2 489	0 307	0 795	0 025	−0 029		
SSR = 019	(−2 67)	(−0 23)	(−3 04)	(−0 21)	(−3 71)	(1 65)	(3 47)	(1 16)	(−4 09)		
$\hat{\rho}$ = .048											
(4) 1935 69	−6 67	−0 491	−0 410	0 055	−2 081	0 545	1.145	0 039	−0 033	−0 112	
SSR = 045	(−1 89)	(−0 69)	(−3 41)	(1 03)	(−2 60)	(2 00)	(3 02)	(1 26)	(−2 98)	(−2 00)	
$\hat{\rho}$ = 524											
(5) 1935 69	−5 572	−0 698	−0 303	−0 020	−1 803	0.595	1 133	0 054	−0.036	−0 069	−0 111
SSR = 036	(−1 70)	(−1 06)	(−3 01)	(−0 41)	(−2 27)	(2 93)	(4 38)	(2 21)	(−4 26)	(−1 36)	(−2 48)
$\hat{\rho}$ = 013											
(6) 1935–69	−3 48	−1 138	−0 179	−0 016	−2 321	0 832	1 097	−0 014	−0 0262		
SSR = 030	(−1 21)	(−2 35)	(−1 66)	(−0 94)	(−3 81)	(4 91)	(3 11)	(−0 56)	(−2 58)		
$\hat{\rho}$ = 745											
(7) 1935–69	−4 698	0 898	−0 425	0 018	−2 009	0 910	1 223	0 041	−0 028		
SSR = 060	(−1 10)	(−1 21)	(−2 95)	(0 73)	(−2 29)	(3 18)	(2 41)	(1 15)	(−1 91)		
$\hat{\rho}$ = 664											
(8) 1935 69	0 178	−0 00040	−0 00054	−0 00106	−0 264	0 202	0 00012	0 00040	−0 0034		
SSR = 00015	(3 25)	(−0 88)	(−3 10)	(−1 26)	(−3 15)	(3 60)	(5 97)	(1 92)	(−6 01)		
$\hat{\rho}$ = 214											
(9) 1938 69	0 188	−0 00077	−0 00075	0 0021	−0 200	0 191	0 000091	0 00038	−0 0023		
SSR = 00014	(2 79)	(−1 38)	(−2 85)	(1 19)	(−2 20)	(2 82)	(3 22)	(1 41)	(2 80)		
$\hat{\rho}$ = 216											

Note:
P_a^0 = clearance rate for murder.
$P_{c/a}^0$ = conviction rate for murder.
$(PXQ_1)_{-1}$ = ratio of current executions to one-year lagged murder convictions.
L = labor force participation rate.
A = fraction of population between 14 and 24 (in equation (6) between 18 and 24 years old).
Y_p = permanent income per capita.
U = unemployment rate.
T = time trend.
P_{63} = zero from 1938–62, $(PXQ_1)_{-1}$ from 1963–69.
C_{63} = zero from 1938–62, 1 from 1963–69.
SSR = sum of squared residuals.
$\hat{\rho}$ = estimate of first-order serial correlation coefficient.

rogates for $P_{e/c}$, the subjective conditional probability of execution given conviction. Neither prior judgement nor the estimated coefficients provide much basis for choosing among variant equations. We chose to concentrate on equation (1) because it is the most common form in Ehrlich's tables and because it permits convenient comparisons with the tradeoffs computed by Ehrlich.

Equation (2), Table 1, shows our attempt to replicate equation (1), based on Ehrlich's (1975b) description of the data sources used. Note that the replication is not precise, though the differences are quantitatively small. The source of the differences probably lies either in minor data collection errors or in the use of different computer programs.[1] While these differences could be reconciled were we to use the same numbers and programs as Ehrlich, we feel that the data behind equation (2) are adequate to examine the validity of Ehrlich's conclusions.[2]

Behind the use of time-series estimation is the assumption that the structure and the coefficients to be estimated remain stable over the sample period. In Ehrlich's case it

[1] We used the instruction *TSCORC* of version 2 7 of *TSP* (Time-Series Processor).

[2] This opinion is based on the fact that our examination focuses on factors which dominate the minor differences between equations (1) and (2) such as the murder rate increase in the 1960's and the logarithmic transformation of the execution rate.

is assumed that the behavior of potential murderers is governed by the same variables with the same coefficients over the period 1935-69. If this assumption is in fact not correct, the estimated function would have little use either as an explanation of the causes of murder or of the policy implications of changing the value of an exogenous variable in the structure.

The assumption can be tested.[3] Consider for example the hypothesis that the murder rate function estimated in equation (2) has the same structure from 1935-62 as from 1963-69. Equation (3) with the time-series truncated at 1962 was estimated to test this hypothesis. The F-ratio, computed from the sums of squared residuals in equations (2) and (3) is 6.00, significant even at the 99 percent level. Thus, the hypothesis of structural homogeneity must be rejected. Further, there is nothing special about the shift point chosen for this test. Similar tests were computed[4] for the four possible structural shift points from 1961-64; the F-ratio indicates a significant shift for each of these periods. It is clear therefore that Ehrlich's equation has been estimated over different regimes, a fact which casts considerable doubt on the validity of his estimates.

Since our primary interest is in the deterrent effect of capital punishment, it is important to note that the coefficient of $(PXQ_1)_{-1}$ is very different in equations (2) and (3), turning from -0.062 and significant to -0.008 and insignificant when the sample period is changed. The F-test for the entire equation does not, however, exclude the possibility that changes other than changes in the coefficient of $(PXQ_1)_{-1}$ have generated the statistical significance.

Equations (4) and (5) provide a test of the hypothesis that the coefficient of $(PXQ_1)_{-1}$ is the same over two sample periods under two different sets of assumptions. In equation (4) we add a variable, P_{63}, equal to zero in the years 1938-62 and to $(PXQ_1)_{-1}$ in the years 1963-69. The t-ratio of this variable gives a test of this homogeneity hypothesis, given that all other coefficients are the same over the two periods. The value of the statistic indicates rejection of the hypothesis at the 90 percent level. In equation (5) we add another variable, C_{63}, equal to 0 from 1938-62 and to 1 from 1963-69. The t-ratio of P_{63} in this equation gives a test of the hypothesis of homogeneity of the coefficient of $(PXQ_1)_{-1}$, given that all variables except the intercept are stable over the two periods. The value of this statistic also suggests rejection of homogeneity.

II

Statistical tests for temporal homogeneity in Section I strongly suggest that the coefficients of relevant variables do not remain the same over the 1935-69 period; inclusion of the last few years is vital if one is to infer some deterrent effect for the conditional probability of execution. One casual explanation of this regression result is quite simple. During the 1963-69 period the logarithm of $(PXQ_1)_{-1}$ falls sharply and monotonically (from -0.65 to -3.91, the latter, 2.3 standard deviations below the sample mean) while the logarithm of Q/N increases from its lowest sample value -3.08, to -2.62, the highest sample value. (See Figure 1.) It is possible, of course, that the rise in murder rates is causally linked to the fall

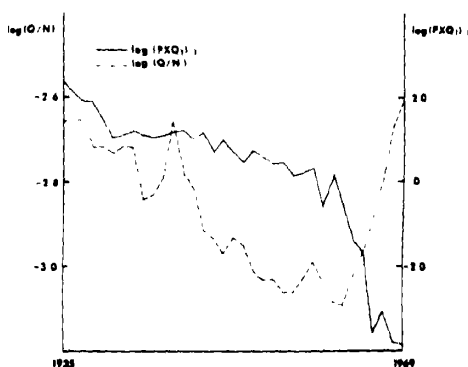



FIGURE 1. EXECUTION RATES AND MURDER RATES, 1935-69

[3] All tests of significance in this paper and in Ehrlich's are based on asymptotic distribution assumptions. Thus what are conventionally called t-ratio or F-ratios in the classical linear regression model will only approximately have the appropriate t or F distribution in these models.

[4] These values are 5.92, 7.07, and 7.36 for the samples ending in 1961, 1963, and 1964, respectively.

in execution rates, as Ehrlich's theory would predict. However the regressions are not in themselves evidence for the existence of such a relationship.

For many observers, alternative explanations of the murder explosion of the 1960's are, a priori, more convincing. As large as increases in murder rates were, the growth rates of other crimes were greater. From 1963 to 1969, per capita reported murders increased 60 percent, robberies by 178 percent, auto theft by 104 percent (see *UCR*, Table 1). Yet few social scientists would choose to connect the increase in auto thefts to the fall in execution rates for murder. Possible nonexclusive explanations for murder rate increases include reduction in the opportunity cost of possessing deadly weapons, racial tension, increases in the difference between economic expectations and opportunities for poor people, reductions in the length of prison sentences. Some of these variables are not easily quantifiable (racial tension, economic frustration, real cost of weapons). Nonetheless their omission from multivariate models of murder rates may so seriously bias the estimated coefficients of included variables that the least squares exercise becomes meaningless. One variable, the expected length of prison sentences given conviction, is not available annually, but is obtainable by state for certain years. It is striking that when the average length of sentences served is included in 1950 and 1960 cross-section regressions, executions lose all explanatory power.[5]

The sensitivity of the estimated coefficient of $(PXQ_1)_{-1}$ to the choice of included variables is notable even for seemingly minor modifications. Ehrlich includes A, the fraction of residential population ages 14–24, to control for shifts in the size of the murder-prone population. Demographic statistics for this age group are easily found in standard reference sources, but the specific choice of the 14–24 year old age group meets no clear theoretical objective. One might plausibly argue on a priori grounds that this age group is unnecessarily broad, young teenagers not being more murder-prone than adults older than 24. Note however that when a narrower target group variable A^*, the fraction of the residential population ages 18–24, excluding armed forces overseas, is substituted for A in Ehrlich's basic equation, the regression results are qualitatively changed.[6] Equation (6) shows the estimated coefficient of $(PXQ_1)_{-1}$ reduced in absolute value and statistically not significantly different from zero.

We do not claim that the major cause of the increase in murders in the 1960's was age group shifts due to the baby boom. This demographic variable may simply be correlated with important omitted variables. What does seem clear, though, is that the data and model employed by Ehrlich do not permit discrimination among numerous plausible explanations of increased murder rates.

Further evidence of sensitivity comes from examining the mathematical transformations of the variables in Ehrlich's model. Theory suggests that an econometric structure should be specified in the mathematical form which conforms most closely to behavioral expectations. In practice, forms are usually chosen which are linear in the parameters to make it easier to interpret the statistical properties of the estimators. And commonly, the logarithmic transformation is chosen to facilitate interpretation of the linear parameter estimates as partial elasticities. In this latter case, the procedure is often justified *ex ante* on theoretical grounds and sometimes *ex post* on statistical grounds. But when the theoretical justification is weak, the results of the estimation are thrown into doubt if they are sensitive to the particular transformation chosen.

In the case of Ehrlich's choice of the logarithmic transformation, the theoretical justification is not convincing. While the elasticity of murder rates with respect to the

[5]See Passell, Tables 1 and 2.

[6]Our source for total residential population is the same as Ehrlich's. The source for the numerator of the fraction was U.S. Bureau of the Census, #114, Table (1933–40), #98, Table 1 (1941–49), #441, #438, Table (1950–69). It was assumed that one-half of all U.S. citizens abroad are 18–24, in order to calculate the remaining resident population for the years 1950–69.

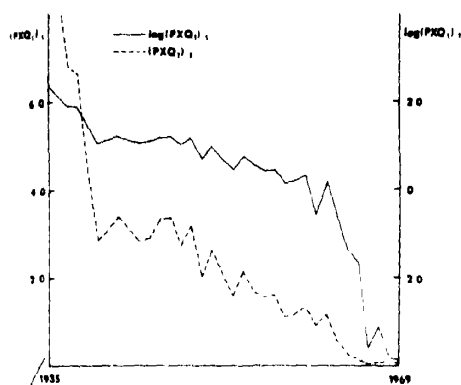



FIGURE 2. EXECUTION RATES, 1935-69, BEFORE AND AFTER *Log* TRANSFORMATION

conditional probability of execution might be approximately constant over some range of the variables, we see no reason why this constant elasticity should hold over the very large range in the observed time-series. However, Ehrlich's finding of significant deterrence effect using 1960's data does appear to be sensitive to the choice of this transformation. Equation (7) is identical to equation (2) with one exception: the data for the conditional probability of execution are employed without transformation to natural *logs*. Note that the coefficient of the execution variables turns positive and statistically insignificant. The intuitive reason for this sensitivity is suggested by the wide difference between the transformed and the untransformed series in Figure 2. Equation (8) uses data for all the variables in untransformed state; the coefficient of execution here is negative, but the t-statistic indicates that it is not significantly different from zero.[7]

Given the apparent sensitivity of the model to these transformations, one might be tempted to use statistical criteria to help one choose the correct transformation. But even if these criteria clearly suggested the choice of a particular transformation, one would still question estimation results which were extremely sensitive to this choice in light of the lack of a strong theoretical justification.[8]

III

In this section we abstract from the criticism of Sections I and II to examine another aspect of the analysis. If one were to accept Ehrlich's estimated structure, would it be reasonable to infer a murder-execution tradeoff from the negative coefficient of $P^0_{e/c}$?

Ehrlich argues that this coefficient ($\hat{\alpha}_3$) should be interpreted as an estimate of the partial elasticity of murder rates with respect to the risk of execution. Thus he is able to derive the estimated marginal tradeoff between executions and murders: $\Delta Q/\Delta E = \hat{\alpha}_3(Q/E)$. If one assumes $\hat{\alpha}_3 = -.065$ and Q and E equal to the time-series mean values, $\Delta Q/\Delta E = -7.7$. For $Q = 10{,}920$ and $E = 41$, $\Delta Q/\Delta E = -17.3$.

It is important to note, however, that behind this interpretation of $\hat{\alpha}_3$ lies an important assumption. The coefficient represents the murder-execution tradeoff only if it is possible to alter PXQ_1 without changing the value of any other independent variable in the equation.[9] Yet social scientists (including Ehrlich) interested in economic models of crime typically acknowledge the simultaneity of the system— hence the logic

[7]Sensitivity to functional form is even more delineated when equation (8) is reestimated without the partially extrapolated deterrence data needed to include the years 1935-37. When equation (8) is estimated from 1938-69 the coefficient for $(PXQ_1)_{-1}$ is positive (.0021) and statistically insignificant with a t-value of 1.19 (see equation (9)).

[8]One statistical procedure for choosing a transformation has been suggested by G. Box and D. Cox, where the choice is reduced to finding a single parameter α in the transformation $x_t = (x_t^\alpha - 1)/\alpha$, of all the variables in the model. As part of our sensitivity analysis we computed the correlation between the actual murder rates (untransformed) and the predicted murder rates (untransformed) as given by an equation fitted to the Box and Cox transformed variables for 100 values of α between 0 and 1. The highest correlation was at .19 which gave a negative and significant coefficient for the execution variable. However, since the model is not stable over the sample period these estimates provide little insight into the problem.

[9]This may not even be possible if one algebraically interprets the deterrent variables as ratios. If one formally differentiates murder rates with respect to executions in Ehrlich's equation (where Q appears in the denominator of the arrest rate) one finds $\partial Q/\partial E > 0$. For further interpretation of this result see the authors.

of calling criminal behavior functions *supply* functions. For the purposes of estimation, Ehrlich explicitly argues that the probability of arrest and conditional probability of conviction are endogenously determined. Thus one must at the very least account for possible adjustment in these two variables before computing the policy implications of changing the probability of execution.

Traditional methodology offers two alternatives for generating an estimate of the tradeoff. One might estimate the rest of the structural equations in the simultaneous system, or one could estimate the reduced form equation for murder rates. The first alternative is beyond the empirical scope of Ehrlich's paper; no attempt is made to specify the other equation in the system. Ehrlich acknowledges the relevance of the second alternative and does produce negative and statistically significant estimates of the reduced form coefficients of the execution variable for certain specifications of $P_{e/c}$ over the period 1934–69. However our confidence in the value of these estimates is limited (a) by the possible sensitivity to functional form and time period of the estimates, and (b) by the untested assumption that the "modified reduced form" chosen by Ehrlich is not biased by omitted variables.

While it may not be possible to calculate the reduced form murder-execution tradeoff, it is interesting to note the potential sensitivity of the sign of that tradeoff to the elasticity of $P_{c/a}$ with respect to $P_{e/c}$. Many legal experts and social scientists believe that increases in $P_{e/c}$ will reduce $P_{c/a}$, since juries and judges will apply stricter standards for convictions when there is a greater prospect of execution.[10] This widely accepted hypothesis was used successfully as an argument by nineteenth and twentieth century legal reform movement leaders bent on abolishing whole categories of capital crimes and granting juries greater discretion in penalties imposed.[11] By Ehrlich's estimated equation (3) in Table 3, if the absolute value of the elasticity of $P_{c/a}$ with respect to $P_{e/c}$, ϵ_{ce}, were greater than .174, the net impact of an increase in execution probabilities would, perversely, raise murder rates. More generally, the impact will be perverse if $|\epsilon_{ce}| > \alpha_3/\alpha_2$.

[10] See for example, Neil Vidmar; Hugo Bedau; James Bennett; Harry Kalven and Hans Zeisel, Zeisel; Robert Knowlton.

[11] See, for example, Maynard Shipley, John McCloskey.

IV. Conclusion

This examination of Ehrlich's estimates has limited focus. One might wish to examine further the theoretical basis for including or excluding variables in the model, the consistency between the data and alternative hypothetical models, the quality of the data (particularly for the deterrence variables P_a^0, $P_{c/a}^0$, $P_{e/c}^0$) used in the test, and the aggregation problems imposed by the use of national data.

We have essentially confined ourselves to a narrower analysis. First, we have shown that Ehrlich's model does not satisfy the statistical requirement of temporal homogeneity and that the results are sensitive to specification of the variables and transformation of the data. This sensitivity raises grave (and in our own opinion, overwhelming) doubt about the utility of Ehrlich's time-series estimates as partial elasticities. Second, we have argued that even if Ehrlich has captured the essence of the murder rate function in his regression, it is not possible to infer from it that a change in legal institutions which increased $P_{e/c}$ would reduce murder rates. In sum, on the basis of Ehrlich's research, it is prudent neither to accept nor reject the hypothesis that capital punishment deters murder.

REFERENCES

Hugo Bedau, *The Death Penalty in America*, New York 1967.

J. Bennett, "A Historic Move: Delaware Abolishes Capital Punishment," *Amer Bar Assn. J.*, Nov. 1958, *44*, 1053–54.

G. Box and D. Cox, "An Analysis of Transformations," *J. Royal Statist. Soc.*, Series B, Jan. 1964, *26*, 211–43.

Robert Dann, "The Deterrent Effect of Capital Punishment," *Friends Social Service*

Series Bull. No. 29, Philadelphia, Mar. 1935.

I. **Ehrlich,** (1975a) "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *Amer. Econ. Rev.*, June 1975, *55*, 397–417.

——, (1975b) "The Deterrent Effect of Capital Punishment: A Question of Life and Death—Sources of Data," unpublished Appendix, May 1975.

Harry Kalven and Hans Zeisel, *The American Jury*, Boston 1966.

R. **Knowlton,** "Problems of Jury Discretion in Capital Cases," *Univ. Pennsylvania Law Rev.*, June 1953, *101*, 1099–1136

J. **McCloskey,** "A Review of the Literature Contrasting Mandatory and Discretionary Systems of Sentencing in Capital Cases," unpublished study for the Pennsylvania Governor's Study Commission on the Death Penalty, Harrisburg 1973.

P. **Passell,** "The Deterrent Effect of the Death Penalty: A Statistical Test," *Stanford Law Rev.*, Nov 1975, *28*, 61–80.

—— **and J. B. Taylor,** "The Deterrence Controversy: A Reconsideration of the Time Series Evidence," in Hugo Bedau and Chester Pierce, eds., *Capital Punishment in the United States*, New York 1976.

Thorsten Sellin, *The Death Penalty*, Philadelphia 1959.

M. **Shipley,** "Does Capital Punishment Prevent Convictions?," *Amer. Legal Rev.*, May-June 1909, *43*, 321–34.

N. **Vidmar,** "Effects of Decision Alternatives on the Verdicts and Social Perception of Simulated Jurors," *J. Personality Soc. Psychol.*, May 1972, *22*, 211–18.

H. **Zeisel,** "Some Data on Jury Attitudes Toward Capital Punishment," unpub. paper, Center for Studies in Criminal Justice, Univ. Chicago Law School 1968.

U.S. Bureau of the Census, *Current Population Reports*. Series P-25, No. 98, Aug. 13, 1954; 114, Apr. 27, 1955; 438, Jan. 21, 1970; 441, Mar. 19, 1970.

U.S. Department of Justice, Federal Bureau of Investigation, *Uniform Crime Report* (*UCR*), Washington.

The Deterrent Effect of Capital Punishment: Reply

By ISAAC EHRLICH*

Despite its title, the comment by Peter Passell and John Taylor does not offer an alternative theory of criminal behavior or any coherent approach that can explain the data they have examined. Moreover, Passell and Taylor (hereafter called P-T) neither challenge the theory nor the statistical methodology I used in my (1975a) paper on the deterrent effect of capital punishment. Indeed, their work is not entirely novel since it duplicates in large measure an earlier published study by W. J. Bowers and G. L. Pierce, to which I have already responded elsewhere (1975b). Their main point lies in emphasizing some apparently negative results that they obtain upon shortening the sample period and altering the regression format. As I hope to show, however, Passell and Taylor have not adequately analyzed the data and their negative findings, and some of the inferences they draw are based on test statistics whose properties are not known. That they do not duplicate the results I reported in my (1975a) study with sufficient accuracy is bothersome in the context of their subsequent regression analysis, especially since others have reproduced my basic results to rounding errors.[1] Finally, Section III of their comment in which they argue for policy implications against reimposition of the death penalty on grounds of an automatic negative association between execution and conviction risks amounts only to speculation. In Sections I–III of this reply I shall address the technical aspects of their critique and their discussion of policy implications. In Section IV I shall report some additional findings concerning the issue of deterrence based upon independent bodies of data and briefly indicate their relation to the results obtained through the time-series analysis.

I

The apparent sensitivity of some of the regression coefficients to deletions of observations relating to both the 1960's and the late 1930's has already been discussed in my own work (1973, n.28 and 1975b, part 3). But while Passell and Taylor have focused their attention on those subperiods in which specific regression results are relatively weak, they have not proceeded to analyze some of the inherent limitations of the data which account for these results.

In the first place, their main test of stability of the regression coefficients across the subperiods 1935–62 and 1963–69 is not defended. The reader may note that the estimation method I used, and which Passell and Taylor adopt, is the three-round nonlinear estimation procedure proposed by Ray C. Fair. To my knowledge, the straightforward application of Chow's test by P-T to test the hypothesis of equality of regression coefficients[2] has not been established as a valid test procedure in the context of this non-linear simultaneous equation procedure. Thus, the properties of their test statistic are not known.

Second, the introduction of a dummy variable distinguishing the 1963–69 subperiod from the earlier period—the one dummy variable which appears to have a pronounced effect in their analysis—does not have an unambiguous effect on the coefficients associated with alternative risk measures. I have found, for example, that when a reduced form estimate of the conditional execution risk—$\widehat{PXQ}_1$ in equation (6)

*University of Chicago and National Bureau of Economic Research. I would like to thank Ronald Gallant, Lawrence Fisher, Randall Mark, and William Wecker for useful suggestions and assistance.

[1] See Brian Forst, Victor Filatov, and Lawrence Klein.

[2] The reader may note that P-T apply the specific variant of Chow's test for the case where the number of degrees of freedom is inadequate for the regular test. They cannot estimate the regression parameters separately for the period 1963–69.

of Table 3 in my (1975a) paper—is used as a regressor, the introduction of that dummy variable (D_{6369}) far from weakening the effect of $\widehat{PXQ}_1$ actually makes it stronger.[3]

What Passell and Taylor fail to note in their analysis are some important differences between subperiods ending in the early 1960's and the full observation set that can explain the apparent changes in coefficients associated with a few of the explanatory variables. Whereas the modified rates of change in the execution risk measures have been stable over the two decades preceding the 1960's, the objective risk of execution declined quite sharply starting about 1960. Variability in deterrence variables used as regressors is particularly small between the late 1930's and the early 1960's.[4] In addition, key variables in the analysis—the murder rate and the conditional risk of execution measures—are highly trended with time over subperiods ending in the early 1960's but much less so over the entire sample period.[5] Regressions based on samples ending in the early 1960's thus are unlikely to produce efficient estimates of the distinct partial association between the murder rate and execution risk.

As implied by the earlier discussion, however, it is not clear that the observed changes in regression estimates across a few subperiods are statistically meaningful. One factor that may contribute to some difference in the magnitude of residuals is that measures of execution risk in 1968 and 1969 are imprecise (see my 1975a paper, Table 2, note b). Indeed, estimates of the negative effect of execution risk based on regressions with 1967 and 1966 as ending dates, although virtually identical to results for the full period, are found to be associated with smaller standard errors.[6] Another such factor is that data concerning national murder statistics in earlier periods were reestimated by the FBI on the basis of the more complete censuses conducted in the 1960's. In addition, the variable *XPOL* (real per capita expenditure on police) is measured for fiscal years since 1961, whereas in the previous years it is defined for calendar years.

But quite apart from the question of the validity of P-T's tests the relevant question is: Given the regressions based upon the full sample and all the various regressions obtained after deletions of specific subperiods, do the results lean in favor of the deterrence hypothesis or do they support the "alternative view" of P-T? I believe the answer is decisively in favor of the former proposition. The results I reported are consistent with sharp predictions concerning the signs as well as the relative ranking of deterrence and other variables. Moreover, the standard errors of the coefficients estimated from the subperiods with earlier ending dates—even those reported by Passell and Taylor—are

[3]The following estimated regression equation relates to the effective sample period 1935–69. All variables are defined in my (1975a) paper. $\hat{\beta}/S_{\hat{\beta}}$ are in parentheses.

$$\ln q^0 = \underset{(-1.30)}{-5.191} - \underset{(-1.50)}{1.087}\Delta^* \hat{P}^0 a - \underset{(-3.36)}{0.392}\Delta^* \hat{P}^0 c \mid a$$

$$\underset{(-4.35)}{-0.155}\Delta^* \widehat{PXQ_1} - \underset{(-1.93)}{2.016}\Delta^* L + \underset{(1.29)}{0.343}\Delta^* A$$

$$+ \underset{(4.83)}{1.344}\Delta^* Y_p + \underset{(1.84)}{0.052}\Delta^* U - \underset{(-6.00)}{0.053}\Delta^* T - \underset{(-3.71)}{0.256} D_{6369}$$

$$\hat{\rho} = -0.26; \quad \hat{\sigma}_e = 0.054; \quad D.W. = 2.17$$

[4]For example, the variance of the modified rates of change of $PXQ_{1_{-1}}$ calculated for the value of $\hat{\rho} = 0.077$ for the period 1935–69 (2.331) is more than 10 times higher than the corresponding variance associated with this variable over the period 1938–62 (.192). At the same time the variances in corresponding values of the rates of change in P^0a and $P^0c \mid a$ fall from .000899 to .000563 and from 0.0199 to 0.00432, respectively.

[5]For example, the zero-order correlation coefficients between these modified rates of change in q^0 and $PXQ_{1_{-1}}$ on the one hand, and the time trend variable on the other, are found to be −.92 and −.92 over the subperiod 1938–62. The corresponding correlations over the full sample period 1935–69 are −.52 and −.46, respectively.

[6]Compare the equation below, for the effective sample period 1935–67, with equation (3) in Table 3 of my (1975a) paper.

$$\ln q^0 = -5.084 - \underset{(-1.73)}{0.853}\Delta^* P^0 a - \underset{(-3.35)}{0.295}\Delta^* P^0 c \mid a \quad (-1.48)$$

$$-\underset{(3.73)}{0.0062}\Delta^* PXQ_{1_{-1}} - \underset{(-2.34)}{1.764}\Delta^* L + \underset{(2.01)}{0.384}\Delta^* A$$

$$+ \underset{(4.52)}{1.105}\Delta^* Y_p + \underset{(2.21)}{0.051}\Delta^* U - \underset{(-6.47)}{0.041}\Delta^* T$$

$$\hat{\rho} = -0.023; \quad \hat{\sigma}_e = 0.039; \quad D.W. = 1.88$$

considerably larger than the corresponding estimates based on the full observation set with 1969 or 1967 as ending dates.[7] The subperiod estimation involves loss of precious degrees of freedom already in short supply. With a sufficiently small number of degrees of freedom one is not likely to reject any particular hypothesis.

II

As I argued in my article (1975b, pp 218-19), and explained further in my subsequent paper (1977) on capital punishment and deterrence, there are both theoretical and statistical reasons for preferring a log-linear specification to a linear format in the variables' natural values. Passell and Taylor's analysis in this regard is inadequate. But the relative superiority of these two forms can be tested systematically via the Box and Cox procedure referred to in my (1975b) piece. Curiously, P-T did attempt to pursue this approach in their present paper and their application of the Box and Cox procedure, if valid, indicates that the optimal transformation they select is much closer to the logarithmic specification ($\lambda = 0$, where λ denotes the coefficient of transformation) than to the linear specification ($\lambda = 1$). Moreover, their sketchy report in footnote 8 of their paper also indicates that at their selected optimal transformation the execution risk variable has a negative and "significant" effect. Their attempt to dismiss these results "since the model is [anyway] not stable over the sample period" is surprising in view of the importance they attribute to their experiments with the simple linear specification where they do not question their results on grounds of instability. More important, as my preceding analysis shows, they have not established the existence of instability.

The application of Box and Cox's analysis of transformations in the context of the estimation procedure used in my time-series analysis is not straightforward. I have attempted to shed light on the issue of optimal transformations via independent cross-sectional regression analyses where data exigencies have dictated employment of classical least squares techniques. The results are decisive. With no exception, the tests conducted on the basis of data from 1940 and 1950 show that under the assumptions of the regression model the simple linear form must be rejected as an optimal transformation within alternative single parameter classes of "power transformations." In contrast, in all of the tests performed, the logarithmic specification cannot be rejected as an optimal transformation at conventional significance levels (see the author 1977). The results of the statistical tests thus lend support to the emphasis I have placed on the log-linear specification in both my time-series analysis of murder and my previous research (1974).

But although the simple linear specification is found to be an inferior format, the qualitative deterrent effects of apprehension, conviction, and execution risks have been observed in my time-series analysis even upon the use of this specification. The results have already been illustrated in my article (1975b, p. 219). I might add that the ranking of the elasticities of the three deterrence variables that are inferred from those results is the same as the ranking predicted by the theory.

III

In their Section III Passell and Taylor speculate that even if my results were valid, an increase in execution risk might raise the murder rate, if the risk of execution is sufficiently negatively related to the risk of conviction. Since their analysis in this regard is not directed against any of my own recommendations, I take it to be their independent contribution to the development of policy implications. Unfortunately, their discussion is not based on systematic theoretical or empirical analyses. P-T defend the proposition concerning a negative association between execution and conviction risks by reliance on "legal experts" and

[7]For example, I have estimated the standard error of the regression coefficient associated with PXQ_2 in the subperiod 1935-63 to be 0.041 as against 0.019 in the subperiod 1935-69 and 0.016 in the subperiod 1935-67.

"legal reform movement leaders" from the nineteenth and the twentieth centuries. We are not told, however, on what scientific grounds these opinions are based.

The "positive" analysis of optimal social defense against murder contained in my (1975a) paper does offer some implications in connection with the relationship between execution and conviction risks (see also the mathematical appendix to my 1977 article) but these implications hardly amount to any automatic rules. The theoretical analysis shows that execution and conviction risks may act as "substitutes," for example, in response to changes in exogenous factors such as *unwarranted* administrative or judicial edicts that affect the frequency of enforcement of the death penalty. But the risks of execution and conviction could also move in the same direction as a consequence of a change in the perceived risk of victimization from murder and related crime. For example, if there were universal agreement among all law enforcement agencies that the rate of crime exceeds what society should bear given its opportunities, then, with appropriate resource expenditures to produce socially optimal magnitudes of deterrence variables, there would be no reason to expect movements in execution risk to generate opposite movements in conviction risk.

I have provided some evidence bearing on the empirical significance of the association between execution and conviction risks for related movements in the rate of murder during the period of my time-series sample (see my 1975a paper, p. 415). It is inconsistent with P-T's predictions. Furthermore, the assertion that execution and conviction risks are negatively associated as a rule is clearly contradicted by the United States experience of the last 15 years. Although the probability of execution decreased dramatically between 1960 and 1968 or thereafter, compensatory movements in probabilities of conviction or arrest are not observed. On the contrary, the estimated unconditional risk of conviction actually decreased from 39 percent in 1960 to about 31 percent in 1969.

As I have repeatedly emphasized, even if effective as a deterrent, capital punishment may not be socially desirable. But the allegation that reintroduction of the punishment necessarily leads to a rise in murder cannot be defended.[8]

IV

My paper (1975a, p. 416) cautioned that the empirical absence of a theoretically important variable—the severity of imprisonment for murder—may have affected the results obtained. Critics of my work also have speculated about potential biases due to absence of a wide gamut of additional factors ranging from latent effects of the baby boom to the collapse of societal values that allegedly account for the results. While the possibility of bias due to omitted variables never can be denied, some of the critics' arguments can be tested. Passell and Taylor, like Bowers and Pierce, argue that the observed negative association between the murder rate and the risk of execution is an artifact of the 1960's—a period in which the rates of other crimes increased even more sharply.[9] If the effect of $P^0e \mid c$ were an artifact of the 1960's however, it should have also affected the trends in other crimes—even crimes against property, which are not expected to be sensitive to

[8]Passell and Taylor carried this argument *ad absurdum* through an exercise (see their fn. 9) where they argue that if society fixed arrests at a constant level, then an increase in executions necessarily will increase murders even though the risk of execution restrains the murder rate. Their analysis is internally inconsistent and their inferences are erroneous. The point is elaborated in the author and Joel Gibbons.

[9]The mere fact that the rate of increase in robbery and other property crimes slightly exceeded that of murder between 1963-69 (68 percent as against 62 percent; Passell and Taylor's quoted numbers are inaccurate) but fell short of it between 1963-73, provides no systematic evidence for the position of the critics, contrary to their assertion, because movements in each crime reflect changes in arrest and conviction risks and other variables that are specific to that crime. Furthermore, P-T ignore the fact that from the late 1930's to 1963 the murder rate in the United States has been continuously on the *decline* while the opposite trend is observed in connection with other felonies. Against this background of conflicting long-term trends in murder and other crimes, the increase in the murder rate since 1964 is even more significant.

TABLE 1 · MURDER SUPPLY FUNCTIONS, EXECUTING STATES *GLS* ESTIMATES[a]
($\hat{\beta}/S_{\hat{\beta}}$ in parentheses)

	Sample	NOB	"R^2"	C (Constant)	T	P^0c	$P^0e \mid c$: $PX4Q$	$P^0e \mid c$: $PX5$	NW	AGE	URB	D_{40}
(1)	1940	33	.9529	8.93	−0.206	−0.709	−0.382		0.443	−2.051	−0.799	
				(2 96)	(−1.70)	(−4.90)	(−3.65)		(6.29)	(−2.29)	(−4.68)	
(2)	1940	33	.9512	10.34	−0 269	−0.678		−0.339	0 485	−2.372	−0.814	
				(3.39)	(−2.17)	(−4 70)		(−3.45)	(7 14)	(−2.57)	(−4.69)	
(3)	1950	34	.9281	3.85	−0.501	−0.765	−0.303		0.488	−0.215	−0.601	
				(1.42)	(−3.88)	(−5 18)	(−4.65)		(6.57)	(−0.30)	(−2.76)	
(4)	1950	35	9473	4 46	−0.574	−0.794		−0.353	0 436	−0.071	−0.769	
				(1 94)	(−5.27)	(−6.31)		(−6.24)	(6.92)	(−0.12)	(−4.20)	
(5)	Pooled	67	.9279	3 85	−0.394	−0 700	−0 311		0 453	−0.347	−0.607	0.469
				(2 04)	(−4 31)	(−6.98)	(−5 94)		(8 95)	(−0.64)	(−4.36)	(3 31)
(6)	Pooled	68	9341	4.06	−0 461	−0.727		−0 334	0 432	−0 216	−0.674	0 446
				(2.26)	(−5.30)	(−7.55)		(−6 59)	(8 89)	(−0.42)	(−5.11)	(3 32)

[a]All variables except D_{40} are measured in natural logarithms. All are weighted by the square root of the urban population (see fn 12) See text for variable definitions Computations performed via the Econometric Software Package The "R^2" reported are for the transformed regression equations with the weighted variables.

changes in execution risk. Preliminary research into the trend of crime in the United States based on a similar econometric specification and time period shows that for all offenses as a group and for crimes against property, changes in the crime rate are insensitive to changes in the execution risk measures. The partial correlation between the murder rate and execution risk does not appear to be artifactual.

Moreover, an independent test of the basic propositions of the model and an investigation of the effect of additional variables are included in my 1977 study based on cross-state regression analyses for the years 1940 and 1950 These years are chosen because the level of enforcement of capital punishment in individual executing states appears to have been sufficiently high and variable to permit a meaningful test of the hypothesis of no deterrence. And the general uniformity in the reporting of urban crime rates (the dependent variable) in the FBI's *Uniform Crime Reports* permits some tests of temporal and cross-sectional homogeneity.

A full analysis of the data and the regression results is included in my 1977 article; Table 1 includes a few illustrations derived from available data on states that had executions. In the equations reported[10] explanatory variables are selected by the availability and compatibility of relevant data in 1940 and 1950. The term P^0c is a measure of the unconditional probability of conviction of murder: the ratio of all commitments to state prisons for murder to an estimate of the total number of murders known in a given state. The term $PX4Q$ is a measure of the conditional probability of execution given conviction, based on the average number of executions for murder in the four years preceding and including the sample year, and $PX5$ is an alternative measure based on the mean number of all executions over a five-year period. The T denotes the median time served in state prisons for murder by prisoners released in 1951. Because of data exigencies it is used as a measure of anticipated length of imprisonment in both 1940 and 1950. The terms NW, AGE, and URB denote, respectively, the percentages of nonwhites, the age group 15–24, and the urban population in the state population. They are introduced partly to serve as "correctors" for relevant variables. The term D_{40} is a dummy variable distinguishing observa-

[10]The regressions are estimated via weighted ordinary least squares. Tests for homoscedasticity led to weighting by the square root of the urban population to generate generalized least squares estimates

tions from 1940 (1) and 1950 (0) in the pooled regressions. It is introduced to account for potential data differences and the use of the variable T in both 1940 and 1950, as well as for the effect of missing variables, such as the income distribution and changes in medical technology (see the author, 1975a, p. 407) which in the time-series analysis may have been accounted for by the trend variable.[11] The general investigation also has examined the impact of the level and distribution of income in the population, the effects of unemployment and labor force participation rates, the significance of alternative measures of the age distribution, and the effect of the risk of death through police intervention. The investigation has also dealt with interdependencies between murder and other related crimes. While many of these additional factors appear to be relevant, the reported effects of the basic set of deterrence and demographic variables introduced in Table 1 is found to be essentially unaffected by their exclusion.

The few reported results speak for themselves. The introduction of imprisonment severity aids in the estimation of deterrence effects. The impact of conviction risk is compatible with the one reported in the time-series analysis.[12] The fact that the elasticity associated with the conditional execution risk measures appears to be higher in absolute magnitude is not incompatible with the proposition that the execution risk effect estimated from data for execution states *only* will be higher than that estimated from aggregate national data (see the author, 1975a, p. 408). As predicted by the theory, the estimated elasticity with respect to conviction risk exceeds that with respect to execution risk. The basic results are found to be stable over time and across regions.

Despite this evidence I do not claim that my general investigation has proven the validity of the general deterrence hypothesis definitively. Many difficulties in the theory's empirical implementation remain and these must be addressed in future work. However, the observed compatibility between the reported time-series and cross-section results, not always present in applications of economic theory, accords further support to the economic approach to criminal activity.

[11] The exclusion of D_{40} from equations (5) and (6) in Table 1 here does not alter the effects of deterrence or any demographic variables other than AGE. Evidently, the latter is systematically higher in 1940 than in 1950.

[12] The regression results reported are not strictly comparable since they are obtained through different estimation procedures. In particular, the GLS estimates may not be consistent. However, time-series results derived through application of OLS techniques for serially correlated errors are quite consistent with the results derived through Fair's simultaneous equation procedure. The following equation has been estimated through the simple Cochrane-Orcutt iterative procedure for the effective period 1935–69. It is comparable to equation (3) of Table 3 in my (1975a) paper:

$$\Delta^* \ln q^0 = \underset{(-1.58)}{-4.541} - \underset{(-1.96)}{1.033}\Delta^* P^0 a - \underset{(-3.58)}{0.292}\Delta^* P^0 c \mid a$$
$$- \underset{(-3.83)}{0.071}\Delta^* PXQ_{1_{-1}} - \underset{(-1.52)}{1.203}\Delta^* L + \underset{(2.45)}{0.513}\Delta^* A$$
$$+ \underset{(4.91)}{1.228}\Delta^* Y_p + \underset{(2.68)}{0.068}\Delta^* U - \underset{(-6.88)}{0.045}\Delta^* T$$
$$\hat{\rho} = 0.023,\ \hat{\sigma}_e = 0.043;\ D.W. = 1.73$$

REFERENCES

W. J. **Bowers and G. L. Pierce,** "The Illusion of Deterrence in Isaac Ehrlich's Research on Capital Punishment," *Yale Law J.*, Dec. 1975, *85*, 187–208.

G. Box and D. Cox, "An Analysis of Transformations," *J. Royal Statist Soc.*, Series B, Jan. 1964, *26*, 211–43.

I. Ehrlich, "The Deterrent Effect of Capital Punishment: A Question of Life and Death," Nat. Bur. Econ. Res. working pap., series no. 18, 1973.

———, "Participation in Illegitimate Activities: An Economic Analysis," in Gary S. Becker and William M. Landes, eds., *Essays in the Economics of Crime and Punishment*, New York 1974, 68–134.

———, (1975a) "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *Amer. Econ. Rev.*, June 1975, *65*, 397–417.

———, (1975b) "Deterrence: Evidence and Inference," *Yale Law J.*, Dec. 1975, *85*, 209–27.

———, "Capital Punishment and Deter-

rence: Some Further Thoughts and Additional Evidence," *J. Polit. Econ.*, forthcoming, Aug. 1977.

——— **and J. C. Gibbons,** "On the Measurement of the Deterrent Effect of Capital Punishment and the Theory of Deterrence," *J. Legal Stud.*, Jan. 1977, *6*.

B. Forst, V. Filatov, and L. R. Klein, "The Deterrent Effect of Capital Punishment: An Assessment of the Estimates," unpub. paper, 1976.

R. C. Fair, "The Estimation of Simultaneous Equation Models with Lagged Endogenous Variables and First Order Serially Correlated Errors," *Econometrica*, May 1970, *38*, 507-16.

U.S. Department of Justice, Federal Bureau of Investigation, *Uniform Crime Report*, Washington, various years.

The Coase Proposition, Information Constraints, and Long-Run Equilibrium: Comment

By HENRY B. HANSMANN*

In a recent paper in this *Review*, William Schulze and Ralph d'Arge employ a partial equilibrium model of two competitive industries with an externality to analyze the well-known "Coase proposition." In particular, they compare both the short-run and long-run efficiency implications of (a) the unadjusted externality case, (b) a Pigovian tax on the output of the firms generating the external cost (the "emitting" firms), (c) a rule making the emitting firms liable to the "receptor" firms, and (d) a situation in which the emitting firms incur no liability, but firms in the two industries are free to bargain costlessly concerning the externality. The analysis in general is illuminating. However, their conclusions regarding the impact of a liability rule, both in the short run and in the long run, appear to be in error. This is of particular significance because, as the authors point out, it is precisely the impact of a liability rule that has been at the center of the controversy over the Coase proposition.

I. The Short Run

According to Schulze and d'Arge, in the short run a liability rule will lead to the optimum level of output in both industries. Under the assumptions they make, however, the liability rule will in fact lead to higher production in the emitting industry than is socially optimal. The error lies in the derivation of their equation (8), which gives the condition for profit maximization in the emitting industry. To see the mistake, it is helpful to rewrite the second equation in their equations (6), which shows the profit for a representative firm in the emitting industry, and from which (8) is derived, as

$$(1) \quad \pi_2 = P_2 y_2 - C_2(y_2) - n_1 D_1(Q_2)/n_2$$

where y_2 is the output of a representative firm in the emitting industry, Q_2 is the total output of the n_2 firms in the emitting industry, C_2 is the direct cost of producing y_2, and $D_1(Q_2)$ is the cost to each of the n_1 firms in the receptor industry of the externality associated with production level Q_2 in the emitting industry. (This formulation follows the authors in assuming that the amount paid to the receptor industry is divided up equally among the emitting firms.)[1] The first-order condition for profit maximization is then

$$(2) \quad \partial \pi_2 / \partial y_2 = P_2 - C_2' - n_1 D_1' \left[\frac{1}{n_2} \frac{dQ_2}{dy_2} \right] = 0$$

This is the same as their equation (8), except that they assume that $dQ_2/dy_2 = n_2$, and thus that the term in brackets equals unity. That is, their formulation assumes that each firm expects that if it increases output by one unit, so will all other firms in the industry, and total industry output will therefore increase by n_2 units. Since they explicitly assume that the emitting industry is competitive, however, a firm in that industry would behave as if $dQ_2/dy_2 = 1$, or, in other words, as if its decisions had no effect on the behavior of other firms. Thus, we can rewrite (2) as

*Assistant professor of law, University of Pennsylvania. I wish to thank William Brainard for valuable comments.

[1]Alternatively, we could assume that liability is divided up among emitting firms according to their output, so that equation (1) instead appears as

$$\pi_2 = P_2 y_2 - C_2(y_2) - n_1 D_1(Q_2) y_2 / Q_2$$

However, so long as n_2 is large and, as the authors assume, $D'' > 0$ (so that marginal damages exceed average damages), short-run output in the emitting industry under a liability rule will still exceed the social optimum.

$$(2') \quad \partial\pi_2/\partial y_2 = P_2 - C_2' - n_1 D_1'/n_2 = 0$$

This condition will be met for individual firms only when their output—and that for the industry as a whole—exceeds the level corresponding to a social optimum. Indeed, if n_2 is large the short-run equilibrium for the emitting industry will be quite close to that in the unadjusted externality case, in which firms in the emitting industry ignore the impact of the externality altogether. In large part, then, the effect of Schulze and d'Arge's liability rule is not to force internalization of the external costs engendered by an emitting firm, but rather simply to shift the burden of those costs away from the firms in the receptor industry and onto the other firms in the emitting industry.

II. The Long Run

In the long run, the authors assert the liability rule results in an overallocation of resources to both industries. Their conclusion is correct so far as the receptor industry is concerned. In the emitting industry, however, the effect of a liability rule might well be underproduction rather than overproduction.

Schulze and d'Arge's argument is based upon their Figure 2. They correctly note that in long-run equilibrium a representative firm in the emitting industry will operate along an average cost curve given by $AC_2 = C_2/y_2 + n_1 D_1/n_2 y_2$. They argue that this average cost curve will always lie below the curve corresponding to the optimal Pigovian tax because, under their assumptions, average damages (upon which liability payments are based) are always below marginal damages (upon which the optimal tax is based). Therefore, they reason, under a liability rule price will be set too low in the emitting industry, and total output will be too high. Yet, as noted above, they are incorrect in stating that the *marginal* cost curve perceived by an individual firm is given by $C_2' + n_1 D_1'$ rather than by $C_2' + n_1 D_1'/n_2$. Consequently, with a liability rule firms should be expected to produce at an output level exceeding that which corresponds to the lowest point on the long-run average cost curve. Thus simply observing that the minimum on one average cost curve lies below the minimum on the other does not suffice to establish their conclusion.

But more remarkably, while Schulze and d'Arge correctly note that a liability rule will induce entry of firms into the receptor industry beyond the social optimum, they ignore the fact that this increase in n_1 will tend to *raise* both marginal and average cost in the emitting industry by increasing the amount of damages that must be paid. Similarly, they appear to ignore the effect of changes in the number of firms in the emitting industry, n_2, upon the position of the marginal and average cost curves both directly and via the terms D_1 and D_1', both of which depend on n_2.

When all of these factors are accounted for, it is clear that the long-run equilibrium price in the emitting industry could as well be above as below the social optimum, and thus that there could as well be too little as too much production in that industry. In fact, one would expect underproduction to be the typical result. Only if marginal damages exceed average damages by a substantial amount, and demand in the receptor industry is quite price inelastic (so that the reduction in cost and price resulting from receipt of compensation would cause little expansion in that industry, and thus little increase in the liability of the emitting firms), would one expect to find overproduction in the emitting industry under a liability rule.

These points can be illustrated with a simple example. Assume that the cost functions for the two industries are given by

$$C_i = (y_i - 100)^3 + 100^3 \quad i = 1, 2$$
$$D_1 = (n_2 y_2)^2/75$$

and that the demand curves for the products of the two industries are given by

$$P_i = 30{,}000 - n_i y_i/2 \quad i = 1, 2$$

These functions give the conventional U shaped average cost curves for both industries, and satisfy as well all of the other conditions set out by Schulze and d'Arge

The conditions for a social optimum are those given by the authors as (their equations (2)–(5)):

$$(3) \qquad P_1 = C_1'$$

$$(4) \qquad P_2 = C_2' + n_1 D_1'$$

$$(5) \qquad P_1 y_1 = C_1 + D_1$$

$$(6) \qquad P_2 y_2 = C_2 + n_1 y_2 D_1'$$

Using the specific cost and demand functions just given, these four equations can be solved (by means of substitution and iterative estimation) for the four unknowns, yielding the (rounded) values[2] $n_1 = 34.6$, $n_2 = 105$, $y_1 = 194$, and $y_2 = 150$. The optimal total output of the emitting industry is therefore $n_2 y_2 = 15{,}750$.

With a liability rule, the long-run equilibrium will be characterized by the four equations

$$(3') \qquad P_1 = C_1'$$

$$(2') \qquad P_2 = C_2' + n_1 D_1'/n_2$$

$$(5') \qquad P_1 y_1 = C_1$$

$$(6') \qquad P_2 y_2 = C_2 + n_1 D_1/n_2$$

which are the same as those given by the authors except for (2′), which is discussed above. Substituting the specific functional forms assumed here, these equations can again be solved for the four unknowns giving $n_1 = 300$, $n_2 = 23.7$, $y_1 = 150$, and $y_2 = 193$. Thus, with a liability rule, the total output of the emitting industry is $n_2 y_2 = 4{,}581$. Rather than being greater than the optimum output for the emitting industry, as Schulze and d'Arge predict, this is in fact less than one-third of the optimum output as calculated above.

[2]The figures given here and in the liability rule case below suggest that firms in each industry are divisible. Only minor adjustments in the figures are necessary if an integral number of firms is required.

REFERENCES

W. Schulze and R. C. d'Arge, "The Coase Proposition, Information Constraints, and Long-Run Equilibrium," *Amer. Econ. Rev.*, Sept. 1974, *64*, 763–72.

The Coase Proposition, Information Constraints, and Long-Run Equilibrium: Reply

By WILLIAM D. SCHULZE AND RALPH C. D'ARGE*

The central theme of our paper on the Coase proposition was that competitive economic models make specialized assumptions on availability and cost of information and transactions. Our conclusion was that when an externality exists in production, a long-run equilibrium achieved through free entry (which assumes certain information constraints) is fundamentally inconsistent with negotiating a Pareto optimal solution.

Unfortunately, Henry Hansmann in his comment has missed this point on the availability of perfect information which is central both to Ronald Coase's arguments and to our paper which attempts to conform with the assumptions used by Coase.

Following Coase, in the short run if the receptor industry jointly brings suit against members of the emitting industry as a *group*, transforming the situation into one where bilateral negotiations can occur (which effectively maximizes joint profits for the two groups), the court in deciding who is or is not liable can be viewed as simply reallocating the gains from trade which come from the negotiations to one group or the other.[1] Assuming that the number of firms in each industry is fixed i the short run, we then suggested dQ_1/dy_2 · n_2. If a group of firms is jointly held respon sible for their pollution activity it appear indefensible to presume they individuall do not consider the effect of pollutio control on the group while simultaneousl arriving at a mechanism and agreemen for it. To assume that each of the emittin firms believes $dQ_2/dy_2 = 1$ is to assum that they have *imperfect information* sinc in fact, this belief is false.[2] The assump

*Associate professor of economics, University of New Mexico, and professor of economics, University of Wyoming, respectively.

[1]Since, as we have pointed out, maximization of joint profits is equivalent to a negotiated solution with perfect information, regardless of property rights in the short run, we can summarize our argument by defining joint profits for our two industries under a negotiated settlement as:

$$\Pi_J = n_1[P_1 y_1 - C_1(y_1) - D_1(n_2 y_2)] + n_2[P_2 y_2 - C_2(y_2)]$$

Firms in the industries can only adjust individual output, but not regulate the number of firms, so joint profits in negotiated solutions are maximized only over y_2 and y_1, taking n_2 and n_1 as given in the short run. Thus, the first-order conditions for a maximum of Π_J are:

$$\partial \Pi_J / y_2 = n_2[P_2 - C_2' - (n_1 D_1')] = 0$$

and

$$\partial \Pi_J / y_1 = n_1[P_1 - C_1'] = 0$$

or prices are set equal to marginal social costs (n_2 $n_1 \neq 0$ for an interior solution), satisfying two of th conditions for a social optimal, equations (2) and (as defined in our original paper. To determine if negotiated solution can achieve optimality in the lon run, let us assume that such a solution satisfies a optimal allocation of resources. Thus, (4) and (5), th zero-profit conditions of our original paper, must als be satisfied. Substituting these conditions into ou definition of joint profits, Π_J, implies

$$\Pi_J^* = n_2 y_2(n_1 D_1') > 0$$

or joint profits in a negotiated solution at the socia optimum are strictly positive under our assumptions Thus, no matter how property rights are assigned be tween the two industries, firms in at least one of th industries must have positive profits at the socia optimum. Clearly, in long-run equilibrium (wher profits go to zero) firms will (under free entry) be in duced by these excess profits to overpopulate one o both of the industries. Excess profits at the *social* opti mum are, however, precisely equal to collections fron an optimal Pigovian tax. Thus, to achieve long-ru optimality with negotiations one must still tax one o both of the industries depending on the type of liabilit rule imposed. This simplified version of our argumen was presented in 1972 by Schulze.

[2]We wonder why Hansmann, if he is assuming tha firms have imperfect information, does not in effec assume that emitter firms perceive $d[n_1 D_1(n_2 y_2)/n_2]/dy_2 = 0$. In other words, why would a firm ordered b a court to make damage payments as a lump sum where these are calculated on an industry damage basis, perceive *any* impact of their own actions o that payment? This, of course, yields a solutio

tions of perceived perfectly elastic input supply and demand for output for individual competitive firms (or rather that such firms cannot change prices) in no way precludes information concerning *nonmarket interactions* between firms. The theory is mute on this point. Coase in attempting to fill this gap assumed perfect information concerning any externality to all negotiating parties. We chose not to argue with this assumption but rather to demonstrate that information constraints are implicitly assumed in the long run which invalidate the Coase proposition under these circumstances.[3]

Hansmann has simply changed the assumption of perfect information for negotiating parties and unsurprisingly achieves different results both from Coase's short-run analysis and our application of negotiated solutions to the allocation of resources in the long run. It is inconsistent to assume that firms in an industry must negotiate to achieve a reduction in emissions yet not be cognizant of the effects of this negotiation on their respective output levels.

identical to the pure unadjusted externality case with a lump sum transfer of wealth. Note here that we assume, as Hansmann apparently does, that the court enforces damage payments under a liability solution, but firms fail to negotiate because they fail to realize gains from trade remain even after the court action is complete.

[3]The precise nature of the long-run information constraint is apparent from the analysis in fn. 1. Under a negotiated solution for an interindustry externality with the optimal number of firms, excess profits remain. The assumption of free entry has traditionally implied that current excess profit levels cause entry or exit of firms. Thus, potential entrants by joining one or the other of the industries are *unaware* that they will as free riders both to the preexisting negotiations and to the "current" profit level—destabilize any preexisting negotiated solution and destroy excess profitability for all firms, including themselves.

REFERENCES

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

H. B. Hansmann, "The Coase Proposition, Information Constraints, and Long-Run Equilibrium: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 459–61.

W. D. Schulze, "Property Rights, Taxation and Environmental Externalities," in *Economics of Natural Resource Development In the West*, rept. no. 3, WAERC Committee on Natural Resources Development, Oct. 1972.

——— **and R. C. d'Arge,** "The Coase Proposition, Information Constraints, and Long-Run Equilibrium," *Amer. Econ. Rev.*, Sept. 1974, *64*, 763–72.

Nontraded Goods, Factor Market Distortions, and the Gains from Trade: Comment

By HIROSHI ONO*

The March 1975 issue of this *Review* contained a controversy over the welfare implications of tariff imposition, resulting from the analysis in Section III of a paper by Raveendra Batra (1973a). The issue was raised by Murray Kemp and Edward Tower. Although this controversy is very interesting, I do not propose to join it, but rather point out that the analysis used in Section II of Batra's paper is misleading. Under the assumptions of gross substitutes, $dD_3/dP_2 > 0$, and $dX_2/dP_2 > 0$ in his notation, he derives the result that an effect of a change in the terms of trade on social welfare mainly depends on factor-market distortions. My basic objection to his analysis is the treatment of β and λ in his paper. He claims that the size of both β and λ depend solely on factor market distortions, independent of the factor market orderings. This argument crucially relies on the hypothesis that changes in capital, dK_i, and labor, dL_i, due to a change in the terms of trade, have the same signs. For the case of Section I, the terms of trade are fixed, and therefore the capital-labor ratio does not change at all. However, as soon as the terms of trade are allowed to vary in Section II, it is no longer necessarily true that capital and labor will change in the same direction in any industry. I construct a simple example below for the home good industry to illustrate this point. Then using Figure 1, I show how a change in the terms of trade affects the capital-labor ratio in each industry as well as allocations of labor and capital to each industry.

Suppose that the social welfare function is given by the following Bergson family type,

$$U = \sum_{j=1}^{3} \beta_j ln D_j \tag{1}$$

It is well known that demand functions derived from (1), which satisfy gross substitutability, are given by:

$$D_j = \beta_i \frac{Y}{P_j} \qquad (j = 1,2,3) \tag{2}$$

where

$$Y = \sum_{j=1}^{3} p_j X_j$$

I shall make my argument as simple as possible. Consider the following Cobb-Douglas production function in each industry:

$$X_j = L_j^{\gamma_j} K_j^{1-\gamma_j} \quad (0 < \gamma_j < 1) \qquad j = 1,2,3$$

Then from marginal conditions

$$k_i = \alpha_j \omega \frac{1 - \gamma_j}{\gamma_j} \tag{3}$$

Substituting this into (7) and (8) in Batra and taking a relative price ratio of p_2 to p_1 as a function of only ω, we find that

$$\frac{d\omega}{dP_2} = \frac{\omega}{P_2(\gamma_2 - \gamma_1)} \tag{4}$$

Then using (3),

$$\frac{dk_i}{dP_2} = \frac{k_i}{P_2(\gamma_2 - \gamma_1)} \tag{5}$$

In a similar fashion, the assumption of full employment determines the allocation of labor as follows:

$$\frac{dL_1}{dP_2} = \frac{k_3 - k_2}{k_2 - k_1} \cdot \frac{dL_3}{dP_2} + \frac{K}{P_2(\gamma_2 - \gamma_1)(k_2 - k_1)}$$

$$\frac{dL_2}{dP_2} = \frac{k_1 - k_3}{k_2 - k_1} \cdot \frac{dL_3}{dP_2} - \frac{K}{P_2(\gamma_2 - \gamma_1)(k_2 - k_1)}$$

*Hokkaido University, Sapporo, Japan. I would like to thank an anonymous referee for his suggestions, which contributed to the diagrammatic illustration in this paper.

From the equilibrium in the home good market (see (4) in Batra), dL_3/dP_2 is determined as follows:

$$\frac{dL_3}{dP_2} = \frac{L_3}{P_2} \cdot \frac{k}{(\gamma_1 - \gamma_2)(\omega + k)}$$

From (3), the relationship between factor shares and the capital-labor ratios is seen as follows:

$$k_1 - k_2 = \frac{\omega(\gamma_2 - \gamma_1)}{\gamma_1 \gamma_2} \qquad (n)$$

Therefore, $dL_3/dP_2 \gtreqless 0$, depending on $k_1 \lesseqgtr k_2$. Since $dK_3 = L_3 dk_3 + k_3 dL_3$, $dK_3/dP_2 = k_3 L_3/P_2(\gamma_2 - \gamma_1) \cdot \omega/(\omega + k)$, and dL_3 and dK_3 have the opposite signs. The terms dK_3 and dL_3 vary in a nontrivial way due to a change in the terms of trade.

Using James Melvin's technique, effects of a change in the terms of trade can be described in a general fashion. In Figure 1 I have drawn the factor endowment box diagram, where the vertical axis stands for labor and the horizontal axis capital. Let A and B be initial equilibrium points. With O' as the origin for the home good, let X_0^* be some arbitrary quantity at A and define the corresponding isoquant as X_0^*. Then the box $ODAE$ represents labor and capital allocated to the import-competing good and the export good industries. For convenience, O and A denote the origins for the import-competing good and the export good, respectively. In Figure 1 we consider the case where $k_3 < k_1 < k_2$, but this factor intensity ordering is not essential in the following. At B, the balance of trade equilibrium is satisfied (see equation (5) in Batra).

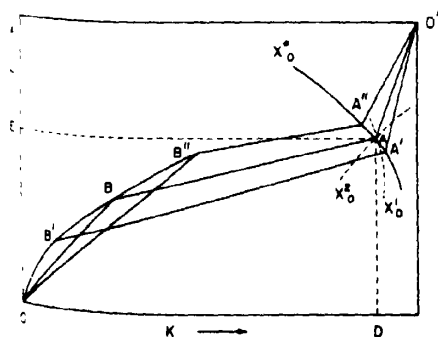



FIGURE 1

Suppose that a change in the terms of trade does not affect the equilibrium production of the home good. Then a new equilibrium lies on the isoquant line X_0^*, along which an increase in one factor necessarily implies a decrease in the other factor. Since factor-market distortions do not disturb the unique functional relationship between the wage-rental ratio and the capital-labor ratio in each industry, whether the economy, due to a change in the terms of trade, moves upward or downward along the X_0^* line depends on whether the wage-rental ratio rises or declines. Suppose that a change in the terms of trade results in a decrease in the wage-rental ratio, which in turn implies a decline in the capital-labor ratio in each industry. A new equilibrium is shown at A' and B'. In the same manner, when the wage-rental ratio rises, A'' and B'' represent new equilibrium points. By considering continuous changes in the terms of trade, in Figure 1 we can draw a locus $OB'BB''$ representing allocation of labor and capital to the export good and the import-competing good industries for various wage-rental ratios. In this example, a decline in the capital-labor ratio in the home good industry comes from the fact that the home good industry hires more labor and less capital.

It is clear from the above argument that we can relax the requirement of factor intensity ordering, and the economy to lie on the same isoquant line. According to the different factor intensity ordering, we can take, say, O' as the origin for the export good. Then the box $ODAE$ shows allocation of labor and capital to the import-competing good and the home good industries. This change does not affect the above argument. Next if a decline in the wage-rental ratio, due to a change in the terms of trade, results in an increase in the production of the home good, the equilibrium locus in the home good industry will be drawn as, say, X_0^1. Corresponding to this locus, we can depict a locus of equilibrium allocations of labor and capital to the export good and the import-competing good industries. On the other hand, if a decline in the wage-rental ratio results in a decrease in the production

of the home good, a new equilibrium point will lie on the locus, say, X_0^2.

REFERENCES

R. Batra, (1973a) "Nontraded Goods, Factor Market Distortions, and the Gains from Trade," *Amer. Econ. Rev.*, Sept. 1973, *63*, 706-13.

——, (1973b) *Studies in the Pure Theory of International Trade*, New York 1973.

——, Nontraded Goods, Factor Mark[illegible] Distortions, and the Gains from Trad[illegible] Reply," *Amer. Econ. Rev.*, Mar. 1975, *6[illegible]* 251-52.

M. C. Kemp and E. Tower, "Nontrad[illegible] Goods, Factor Market Distortions, an[illegible] the Gains from Trade: Comment," *Ame[illegible] Econ. Rev.*, Mar. 1975, *65*, 249-50.

J. R. Melvin, "Production and Trade wit[illegible] Two Factors and Three Goods," *Ame[illegible] Econ. Rev.*, Dec. 1968, *58*, 1249-68.

Nontraded Goods, Factor Market Distortions, and the Gains from Trade: Reply

By RAVEENDRA BATRA*

In 1973 when I wrote the paper dealing with the factor market distortions in the presence of nontraded goods, I had little reason to believe that such an apparently simple problem would turn out to be so devious. The paper has attracted valid criticism first from Murray Kemp and Edward Tower, and now from Hiroshi Ono. In order to settle the matter once and for all, I present a brief and simpler analysis of the same model, one that seems to be infallible. Furthermore, the general theorem obtained applies to all normative questions of factor-market distortions, regardless of the presence and absence of nontraded goods—a feature that was absent in the earlier paper.

THEOREM 1: *In the presence of interindustry wage differentials, the implications of any parametric change in the economy for social welfare depend, among other things, on the employment of labor in the industry paying the higher wage rate.*

I will illustrate this theorem for the case where the parametric change involves an exogenous change in the terms of trade of a small economy.[1] Consider an economy facing a social utility function

$$(1) \qquad U = U(D_1, D_2, D_3)$$

and a budget constraint

$$(2) \qquad D_1 + \frac{P_2}{P_1} D_2 + \frac{P_3}{P_1} D_3 = X_1 + \frac{P_2}{P_1} X_2 + \frac{P_3}{P_1} X_3$$

where D = consumption, X = production, and P = price; and i denotes goods one, two, and three (i = 1, 2, 3) with the first two goods being traded goods and the third good being the nontraded good.

Differentiating (1) and (2) with respect to P_2 and remembering that the marginal rates of substitution reflect price ratios in consumer's equilibrium, we obtain:

$$(3) \qquad \frac{1}{U_1}\frac{dU}{dP_2} = \frac{dX_1}{dP_2} + \frac{P_2}{P_1}\frac{dX_3}{dP_2} + \frac{P_3}{P_1}\frac{dX_3}{dP_2} + \frac{(X_2 - D_2)}{P_1}$$

From now on assume for simplicity that all prices are initially equal to one. If the production functions are given by $X_i = F_i(K_i, L_i)$ with K and L as the capital and labor inputs, respectively, then after differentiation (3) is written as:

$$(4) \qquad \frac{1}{U_1}\frac{dU}{dP_2} = \sum_{i=1}^{3} F_{Li}\frac{dL_i}{dP_2} + \sum_{i=1}^{3} F_{Ki}\frac{dK_i}{dP_2} + X_2 - D_2$$

where F_{Li} and F_{Ki} are respectively the marginal products of labor and capital in the ith sector. In the factor-market equilibrium under perfect competition, $F_{K1} = F_{K2} = F_{K3}$, whereas with full employment of inelastically supplied capital $\Sigma dK_i = 0$. Therefore the second term of (4) reduces to zero. As regards to the first term, $(F_{L1}/\alpha_1) = (F_{L2}/\alpha_2) = (F_{L3}/\alpha_3)$ where $\alpha_i \geq 1$, represents the wage differential.[2] Since $\Sigma dL_i = 0$, then (4) can be written as

$$(5) \qquad \frac{1}{U_1}\frac{dU}{dP_2} = F_{L2}\left[\left(\frac{\alpha_1}{\alpha_2} - 1\right)\frac{dL_1}{dP_2} + \left(\frac{\alpha_3}{\alpha_2} - 1\right)\frac{dL_3}{dP_2}\right] + (X_2 - D_2)$$

This is then a simple equation that determines the implications of a change in the

*Southern Methodist University.

[1] The model used is the same as in my earlier paper.

[2] See the author, p. 707.

foreign price of the second good for social welfare.[3] As can be seen clearly, the welfare change depends on the response of labor utilization to a change in P_2 as well as the amount traded of the second good, which I will suppose is the exported good, so that $X_2 - D_2 > 0$. In the absence of any wage differential $\alpha_i = 1$, so that

$$\frac{1}{U_1}\frac{dU}{dP_2} = X_2 - D_2 > 0$$

which is the standard result. On the other hand, suppose that the wage differential exists, but the nontraded good does not, so that $dL_1 = 0$. Here the results also depend upon the sign of $(1 - \alpha_1/\alpha_2)$ as well as dL_1/dP_2. The question of how labor allocation is influenced by the change in P_2 becomes germane. In the case of "normal" price-output response $dL_1/dP_2 < 0$ so that the result then depends on $(1 - \alpha_1/\alpha_2)$. This is nothing but the result obtained by myself and Prasanta Pattanaik.

Finally suppose that the nontraded good exists and that the differential is paid by industry 1, so that $\alpha_3 = \alpha_2 = 1 < \alpha_1$. Here

$$\frac{1}{U_1}\frac{dU}{dP_2} = F_{l2}(\alpha_1 - 1)\frac{dL_1}{dP_2} + (x_2 - D_2)$$

Since $\alpha_1 > 1$, the change in social welfare depends again on how the labor utilization in the industry paying the differential responds. Similarly, if the differential is paid by the nontraded good, so that $\alpha_1 = \alpha_2 = 1 < \alpha_3$, the welfare change would depend on the labor response in good 3. If on the other hand, good 2 paid the differential, so that $\alpha_1 = \alpha_3 = 1 < \alpha_2$, then since $(1 - \alpha_1/\alpha_2) = (1 - \alpha_3/\alpha_2)$ and since $dL_2 = -(dL_1 + dL_3)$, (5) would become:

$$\frac{1}{U_1}\frac{dU}{dP_2} = -F_{L2}\left(\frac{\alpha_3}{\alpha_2} - 1\right)\frac{dL_2}{dP_2} + (X_2 - D_2)$$

Here again the results would depend on how the employment of labor responds in the industry paying the wage differential.

Until now Theorem 1 has been illustrated by means of an exogenous change in the terms of trade. Similar results can be derived by introducing other parameters such as tariffs and taxes. The appeal of the procedure developed here is not only its simplicity but also its validity, regardless of the presence or absence of nontraded goods.

[3]Equation (5) is quite different from its counterpart, equation (17), obtained in the author, p. 708. The reader can readily appreciate the simplicity of the procedure developed here.

REFERENCES

R. Batra, "Nontraded Goods, Factor Market Distortions, and the Gains from Trade," *Amer. Econ. Rev.*, Sept. 1973, *63*, 706–13.

——— **and P. Pattanaik**, "Domestic Distortions and the Gains from Trade," *Econ. J.*, Sept. 1970, *80*, 638–49.

M. C. Kemp and E. Tower, "Nontraded Goods, Factor Market Distortions, and the Gains from Trade: Comment," *Amer. Econ. Rev.*, Mar. 1975, *65*, 249–51.

H. Ono, "Nontraded Goods, Factor Market Distortions, and the Gains from Trade: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 464–66.

Short-Term Interest Rates as Predictors of Inflation: Comment

By JOHN A. CARLSON*

In a recent article in this *Review*, Eugene Fama concludes that "one ... cannot reject the hypothesis that all variation through time in one- to six-month nominal rates of interest mirrors variation in correctly assessed one- to six-month expected rates of change in purchasing power" (p. 282). Since acceptance (or nonrejection) of this hypothesis over periods as short as a few months is counter to much of today's received doctrine about what influences market rates of interest, it is important to point out that the tests he conducts do not rule out other theoretical models so long as they too are consistent with the observed data.

The nominal rate of interest can be defined as the sum of an expected rate of inflation and an expected real rate of interest. Fama's conclusion then rests on two assumptions: (a) There is a constant expected real rate of interest. (b) All relevant information about future inflation is fully incorporated in the expected-inflation component of the market rate of interest.

Both assumptions will be contradicted by evidence to be presented here. The first assumption does not hold up when an expected-real-rate series is constructed from survey data on inflation expectations. The most remarkable regularity in the series is that the short-term (six months to a year) expected real rate falls during recessions, when the short-term marginal productivity of capital would be expected to fall.

The second assumption will be challenged on Fama's own ground by showing that significant information about subsequent inflation has not been fully reflected in nominal interest rates and by arguing that the common trend in the data gives rise to the statistical illusion that variations in interest rates on Treasury Bills are good predictors of variations in inflation.

I. Behavior of an Expected Real Rate

If a sample were to be taken of individual participants in and observers of financial markets, and they were asked what will happen to prices of goods and services over the term to maturity of a particular Treasury Bill, it is plausible that the average expectation in that sample will tend to move with the expectations of the representative trader. It would then be possible to construct a variable that we may call an expected real rate of interest.

Before turning to the evidence, we should consider how to interpret patterns that might arise. First, as Irving Fisher and many others have argued, we should see a positive association between expected inflation and interest rates. Briefly, suppose on balance that people have just come to believe that inflation will be higher than it is now. In order for those expectations to have any effect on the yields of Treasury Bills, some holders of a bill or some potential buyers now think that a better return can be obtained by somehow buying into items that will appreciate in value at least as fast as the expected inflation rate. This would lead to an increase in supply or a decrease in demand for a Treasury Bill at the current price. Either is postulated to lower the price and raise the yield. Thus, yields should be positively associated with expected inflation rates if the alternative investments are readily available.[1]

The foregoing line of argument, however, also allows other factors affecting supply

*Professor of economics, Purdue University. I wish to thank Pat Hendershott, Bill Dunkelberg, Ben Friedman and Ed Kane for comments on an earlier draft, Jeff Levin and Tommy Stanley for their fine research assistance, and the NSF for financial support. I retain responsibility for the views expressed.

[1]Along this line, Patric Hendershott and James Van Horne find Martin Feldstein and Otto Eckstein's model theoretically deficient because it is missing an asset that can earn a real return.

and demand for Treasury Bills to have an influence on the market yields, at least over relatively short periods, such as a few months. If so, real rates may vary systematically in response to specific stimuli. Standard models postulate, for example, that there will be liquidity effects (see, for example, Milton Friedman). Sizable increases in the money supply may temporarily drive down the nominal rate of interest. Without an immediate matching decrease in expected inflation, the real rate of interest must also fall.[2]

There are other possibilities, but since one pattern does stand out in the data to be presented, we shall move directly to that point. Waves of optimism or pessimism of the sort John Maynard Keynes discusses in conjunction with investment decisions may cause short-run increases or decreases in the *expected* productivity of capital. Those expectations in turn could be reflected in financial markets, such that expected real interest rates fall when investors become pessimistic and rise when they become optimistic about profit opportunities for new capital goods.

Data for computing a time-series of expected real returns exist over a fair length of time. Since the late 1940's, Joseph Livingston, a financial columnist in Philadelphia, has conducted a semiannual survey of business and academic economists who are deemed to be knowledgeable observers of the economy and of financial markets. Most of them have been involved in making economic forecasts and so are likely to have a general common awareness of the information available about the state of the economy at the time of their forecasts. The forecasts in these Livingston surveys include estimates of what the Consumer Price Index (*CPI*) will be six and twelve months beyond the survey month. They are submitted in late November or early December, and in late May or early June.

At the time these forecasts are made, most of the participants in the December surveys know the October *CPI* and those in the June surveys know the April *CPI*. Therefore, the expected rates in inflation reported in Table 1 have been computed as follows. An arithmetic average is taken of the individual forecasts of the *CPI* six and twelve months beyond the survey month. Then, since the latest *CPI* released prior to these forecasts is for two months before the survey month, the implied expected rates of inflation are calculated over the eight and fourteen months between the latest known *CPI* and the *CPI* being predicted. These are expressed at annual rates.[3]

Considering when the questionnaires are returned, the average market yields for the first week ending in the survey month coincides reasonably well with when the price expectations are submitted. These yields are reported in Table 1 together with the week ending dates. Finally, the expected rate of inflation, based on forecasts of the *CPI*, the same index Fama uses, is subtracted from the market yield on securities with maturity dates that correspond roughly to the dates of the price forecasts. The difference is shown in Table 1 as the expected real return.

The patterns are similar for the 6-month and the 12-month returns. Therefore, the 12-month data are plotted in Figure 1 because there are more observations. The 6-month yields were not reported prior to 1959. Inspection of Table 1 and a look at Figure 1 reveal a number of points.

The trend in nominal yields parallels the trend in expected inflation rates, with the result that there is little evident trend in the expected real rates of interest (assuming expected real returns rebound to normal levels after the 1974 dip). This is consistent with the notion that financial markets adjust to the real forces of productivity and thrift over fairly long periods of time and that these forces change only gradually. The ex-

[2]At one point, Fama seems to agree with this when he writes that prior to the Treasury-Federal Reserve Accord of 1951 "a rich and obstinate investor saw to it that Treasury Bill rates did not adjust to predictable changes in inflation rates" (p. 274). If pre-Accord government policy could *peg* interest rates, then surely subsequent policy actions can at least temporarily offset predictable inflation and force variations on the expected real return.

[3]Additional details are available in the author (197-, 1977).

TABLE 1

Year	Week Ending	Yields[a] on 6-Month Bills	Yields[a] on 9 to 12 Month Issues	Expected Inflation Rate[b] Next 6 Months	Expected Inflation Rate[b] Next 12 Months	Expected Real Return Next 6 Months	Expected Real Return Next 12 Months
1953	June 6		2.59	-1.01	-1.50		4.09
	Dec. 5		1.50	-1.16	-1.08		2.58
1954	June 5		.81	- .52	- 11		.92
	Dec. 4		1.01	.15	.05		.96
1955	June 4		1 74	.50	.27		1.47
	Dec. 3		2.44	.81	.54		1.90
1956	June 2		2.74	.35	.50		2.24
	Dec. 1		3.23	1.42	1.08		2.15
1957	June 1		3.42	1.14	1.19		2.23
	Dec. 7		3.33	.07	.23		3.10
1958	June 7		.91	.06	.32		.59
	Dec. 6		3.30	.64	.80		2.50
1959	June 6	3.30	3.99	.62	1.00	2.68	2.99
	Dec. 5	4.86	4 93	.95	1.04	3.91	3.89
1960	June 4	3.18	3.87	45	.70	2.73	3.17
	Dec. 3	2.70	3.05	.19	.64	2.51	2.41
1961	June 3	2.60	2.98	1.01	1.15	1.59	1.83
	Dec. 2	2 78	2.98	1 07	1.21	1.71	1.77
1962	June 2	2.76	3.00	1.02	1 06	1.74	1.94
	Dec. 1	2 94	2 95	.99	1 10	1.95	1.85
1963	June 1	3.06	3.17	1.03	1.05	2.03	2.12
	Dec. 7	3.68	3.76	82	.98	2.86	2.78
1964	June 6	3.57	3.84	1.09	1.24	2.48	2.60
	Dec. 5	3.97	4.04	1.29	1.23	2.68	2.81
1965	June 5	3.92	4.02	.94	1.07	2.98	2.95
	Dec. 4	4.26	4.36	1 57	1.68	2.69	2.68
1966	June 4	4 75	5.00	1.83	2.08	2.92	2.92
	Dec. 3	5.26	5.32	2 06	2.19	3.20	3.13
1967	June 3	3.74	4.12	2.14	2.40	1.60	1.72
	Dec. 2	5.49	5 60	2.66	2.83	2.83	2.77
1968	June 1	5.86	6.20	3.09	3.10	2.77	3.10
	Dec. 7	5.77	5.73	2 72	2.91	3.05	2.82
1969	June 7	6.58	6.78	3.15	3.44	3.43	3.34
	Dec. 6	7.83	8.11	3.55	3.60	4.28	4.51
1970	June 6	6.88	7.52	3.50	3.64	3.38	3.88
	Dec. 5	4.95	5.05	3.56	3.80	1.39	1.25
1971	June 5	4.52	4.99	3.90	4.12	.62	.87
	Dec. 4	4.42	4.63	3.03	3.23	1.39	1.40
1972	June 3	4.20	4.66	3.57	3.80	.63	.86
	Dec. 2	5.18	5.35	3.24	3.48	1.94	1.87
1973	June 2	6.99	7.13	4.11	4.21	2.88	2.92
	Dec. 1	7.77	7.36	5.34	5.36	2.43	2.00
1974	June 1	8.26	8.46	7.12	6.84	1.14	1.62
	Dec 7	7.34	7.65	7.67	7.50	- .33	.15
1975	June 7	5.48	6.15	5.54	5.60	- .06	.55
	Dec. 6	6.04	6.65	5.84	5.85	.20	.80

[a]Source *Federal Reserve Bulletin*, various issues.
[b]Obtained from Livingston's survey of economists.

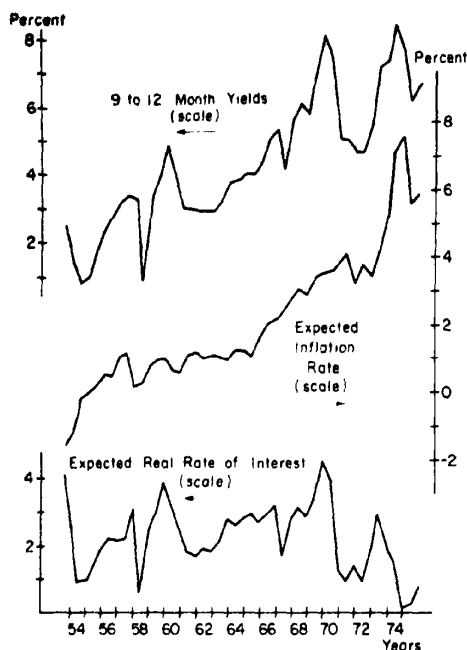



FIGURE 1

pected real returns fluctuate around 2.5 percent, ranging between 0 and 4.5 percent in the years 1953 75. (The 6-month forecasts by late 1974 even indicate a slightly negative expected real return.)

The most notable declines in nominal interest rates occur in 1953 54, 1957–58, 1960, 1966–67, 1969 70, and 1974 75. All of these, with the possible exception of 1966–67, are associated with recessions. Fama's interpretation is that the market, efficiently using information, was anticipating a decline in the rate of inflation. In all of those episodes, however, except in 1960 and in 1975, the expected rate of inflation *rose* according to the Livingston data. In 1960 the expected rate of inflation fell by considerably less than the nominal rate of interest.

In early 1975, the fall in nominal interest rates did mirror a fall in expected inflation. That was also the largest 6-month drop in expected inflation yet observed in the Livingston data. For once the decline in nominal rates was not attributable to a concurrent decline in the expected real rate. Whatever drives down the real rate, however, whether it be expected inflation rising faster than nominal rates or nominal rates falling with little change in expected inflation, the severity of a recession may well be the prime determinant of how low the expected short-term real rate falls.

II. Weakness of the Regression Tests

If expected real rates fluctuate so much why do Fama's tests support the assumption of a constant real rate? I shall argue that two phenomena offer answers: (a) a statistical domination of expected inflation in interest rates adjusting to relatively large differences in inflation in the sample and (b) a systematic underestimation of inflation when expected real rates are most likely to be high.

$$\Delta_t = a_0 + a_1 R_t + e_t \quad (1)$$

where R_t is the interest rate on a Treasury Bill, Δ_t is the actual percentage change in purchasing power over the term to maturity of the bill, and e_t is an error term. (Note that Δ is negative when prices are rising.) If the expected real rate is constant and if correctly anticipated inflation is on average incorporated into the market rate of interest, then the coefficient a_1 should be approximately minus one. Equation 1 in Table 2 is a replication of Fama's estimates for 1-month Treasury Bills.[4] The estimated a_1 coefficient is very close to minus one.

In terms of hypothesis testing, what alternative hypotheses should we reject on the basis of this estimate? It may be that not much has been ruled out. To see this point, consider a simple artificial example with eight periods of observations:

Period	R	Δ
1	4	0
2	4	−4
3	6	−4
4	7	−4
5	8	−4
6	8	−9
7	9	−8
8	10	−7

[4] I am grateful to both Fama and Schwert for allowing me to use these data.

The example has been constructed so that whenever R is unchanged, Δ falls (i.e., the inflation rate rises) and whenever R goes up, Δ stays the same or goes up (inflation is unchanged or lessens). Clearly, changes in Δ are badly predicted by the immediately preceding changes in R, and yet the *OLS* regression equation with these data is:

$$\Delta_t = 2.0 - R_t \quad (2)$$

Taken literally, equation (2) implies that a one-point increase in R results in (or is followed by) a one-point decrease in Δ. In the example, however, the sequence of events is that an increase in inflation is followed with a lag by increases in interest rates. While efficiency is not rejected by the regression estimates, neither should "inefficiency" of forecasts be ruled out.[5] Strong trends in the data present a major empirical problem.

If we turn to the *U.S.* data and select a period in which there is relatively little trend, then the short-term changes in the expected real interest rate and the errors in forecasting inflation a short period ahead may not be swamped statistically by the possibility of interest rates eventually rising in response to higher rates of inflation.[6] In looking at the Livingston data, we see that in the period from late 1956 to mid-1965 the expected inflation rate stayed around 1 percent per year. Prior to that period it had been rising toward 1 and after that it rose markedly above 1. During the period nominal interest rates did have considerable variation.

Equation (1) was therefore estimated for the months from the second Livingston survey in 1956 to right after the first survey in 1965, a total of 105 monthly observations. The estimates are shown as equation 2 in Table 2. The a_1 coefficient is now much closer to zero and by conventional tests of significance, the hypothesis must be rejected that the true coefficient equals minus one.

The same pattern recurs for 2- and 3-month bills. Equations 5 and 9 (Table 2) replicate Fama's corresponding equations. With the 2-month bills, the equation calls for the use of nonoverlapping two-month intervals. This could be done beginning in January 1953 or in February 1953. Either gives approximately the same coefficients. Those beginning with the first month in the sample are reported in all the equations 5 through 12 in Table 2.

The regressions run with the two- and three-month data for the 1956-65 subperiod are numbered 6 and 10. The results are similar to those for 1-month bills (reported as equation 2). The estimates of the coefficient on R_t systematically depart from minus one for every set of nonoverlapping two- or three-month data regressed over the 1956–65 subperiod, not just those reported in Table 2. None of the coefficients comes within two standard errors of minus one.

This rejection of Fama's joint hypothesis in periods in which trend effects are negligible arises not only because the expected real rate varies but also because information about subsequent inflation is not fully incorporated in the interest rate variable. I show this latter point by using another variable that has a good record in predicting inflation. The variable, suggested in a note by Irwin Kellner, is the ratio of seasonally adjusted employment to noninstitutional population over 16 years of age. From an inspection of Kellner's chart, I chose a six-month lag arbitrarily without trying to find a best-fit lag. Thus, the variable denoted

[5] Another of Fama's tests of efficiency is a lack of autocorrelation in the residuals In the example, the residuals have a slight negative autocorrelation but even that could be removed by a somewhat more complicated example in the same spirit as the one presented, with unexpected inflation inducing subsequent changes in interest rates.

[6] In the spirit of the example, I broadened the maintained hypothesis to include current inflation as a function of future interest rates. Let R_{t+i} replace R_t in equation (1) and let $i = 0, 1, \ldots, 5$ in turn With a null hypothesis that $a_1 = 1.0$ and using the data from 1953-71, one cannot reject the null hypothesis at conventional levels of significance with *any* of the values of i that were considered. However, somewhat to my surprise, the fit with $i = 0$ was marginally better than the other cases This is what Fama's model predicts, although the differences in the coefficients relative to their standard errors do not allow any clear-cut statistical claim of superiority for Fama's particular hypothesis Furthermore, it would seem with Fama's model that the fits should get progressively worse as i increases In fact, the fit is better with $i = 4$ than with $i = 1, 2$, or 3.

TABLE 2—REGRESSIONS OF CHANGE IN PURCHASING POWER ON INTEREST RATES R_t AND ON THE RATIO OF EMPLOYMENT TO POPULATION E/P FOR DIFFERENT HORIZONS AND SELECTED PERIODS
(Standard Errors in Parentheses)

Equation	Horizon in Months	Period	Constant[a]	R_t	E/P	R^2	DW
1.	1	1/53-7/71	.069	−.976		.29	1.79
			(.030)	(.102)			
2.	1	11/56-7/65	−.053	−.361		.02	1.95
			(.064)	(.279)			
3.	1	1/53-7/71	−.020	−.639	−.088	.36	1.93
			(.034)	(.119)	(.018)		
4.	1	11/56-7/65	−.042	−.409	−.080	.10	2.07
			(.061)	(.268)	(.026)		
5.	2	1/53-7/71	.159	−.960		.42	1.65
			(.066)	(.107)			
6.	2	11/56-7/65	−.105	−.322		.02	1.74
			(.140)	(.286)			
7.	2	1/53-7/71	.001	−.677	−.153	.50	1.90
			(.073)	(.123)	(.038)		
8.	2	11/56-7/65	−.092	−.350	−.126	.14	2.10
			(.133)	(.272)	(.049)		
9.	3	1/53-7/71	.223	−.908		.48	1.99
			(.105)	(.110)			
10.	3	11/56-7/65	−.205	−.247		.02	2.15
			(.253)	(.329)			
11.	3	1/53-7/71	−.053	−.588	−.258	.59	2.33
			(.113)	(.122)	(.058)		
12.	3	11/56-7/65	−.195	−.260	−.268	.25	2.66
			(.224)	(.292)	(.085)		

[a]Constants are higher than those reported by Fama because Δ and R have been multiplied by 100. The E/P is measured in deviations from its sample mean so that it will not have much effect on the estimate of the constant term.

E/P is the employment/population ratio six months before the month in which the interest rate is determined.

Once again, refer to Table 2 and compare equation 3 with equation 1. When E/P is added to Fama's equation with the one-month data, it enters significantly; with standard normality assumptions, the t-ratio is almost 5. This information has apparently not been fully reflected in the one-month interest rates because the coefficient on the interest-rate variable now departs significantly (by 3 standard errors) from Fama's predicted value of minus one. The results for 2-month and 3-month bills, shown in equations 7 and 11, respectively, tell precisely the same story.[7]

For comparative purposes, equations 4, 8, and 12 in Table 2 show the estimates of the coefficients for the regressions with E/P in the subperiod 1956–65. In that period, at least, the coefficients on E/P are close to those for the entire 1953–71 period. Note also the subperiod coefficients on R_t are not changed by much with the inclusion of E/P.

III. Summary and Conclusion

To conclude, variations in short-term interest rates are not good predictors of variations in inflation rates. Fama's regression tests over periods with substantial trends in both inflation and interest rates cannot rule out a world in which unexpected inflation precedes changes in interest rates. When the regressions are run over periods with little variation in expected inflation, one is led to reject Fama's hypothesis.

Furthermore, both of the key assump-

[7]The progressively higher coefficients (in absolute value) for E/P arise because the dependent variable, i.e., the percentage change in purchasing power, tends to be greater when computed over more months.

tions are of dubious validity. First, evidence has been presented that expected short-term real interest rates do have notable variation. They fall during recessions when the short-term marginal productivity of capital is likely to have fallen. Second, interest rates do not appear to be efficient predictors, if efficiency means that all available information about subsequent inflation rates are incorporated in interest rates.[8] Information about inflation, not fully reflected in interest rates, is provided by an additional variable, the ratio of employment to population.

[8] Fama carefully states that his empirical results support efficiency only in the sense of using all the information available in time-series of past inflation rates. This conclusion is respectfully disputed by Nelson and Schwert. My challenge extends to Fama's broader meaning of efficiency. Patrick Hess and James Bicksler, by contrast, consider results seemingly contrary to efficiency as "very disturbing" and proceed to find other explanations that will reconcile the evidence with a belief in efficient markets.

REFERENCES

J. A. Carlson, "Are Price Expectations Normally Distributed?" *J. Amer. Statist. Assn.*, Dec. 1975, *70*, 749-54.

——, "A Study of Price Forecasts," *Ann. Econ. and Soc. Measurement*, Winter 1977, *6*, 27-56.

E. F. Fama, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269-82.

M. Feldstein and O. Eckstein, "The Fundamental Determinants of the Rate of Interest," *Rev. Econ. Statist.*, Nov. 1970, *52*, 363-75.

Irving Fisher, *The Theory of Interest*, New York 1930.

M. Friedman, "Factors Affecting the Level of Interest Rates, Part I," in Donald P. Jacobs and Richard T. Pratt, eds., *Savings and Residential Financing, 1968 Conference Proceedings*, Chicago 1968, 10-27.

P. H. Hendershott and J. C. Van Horne, "Expected Inflation Implied by Capital Market Rates," *J. Finance*, May 1973, *28*, 301-14.

P. J. Hess and J. L. Bicksler, "Capital Asset Prices Versus Time Series Models as Predictors of Inflation," *J. Finance Econ.*, Dec. 1975, *2*, 341-60.

I. L. Kellner, "The True State of the Labor Market," Manufacturers Hanover Trust Co., Oct. 1975.

John Maynard Keynes, *The General Theory of Employment Interest and Money*, London 1936.

C. R. Nelson and G. W. Schwert, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1977, *67*, 478-86.

Board of Governors of the Federal Reserve System, *Fed. Res. Bull.*, various issues.

Short-Term Interest Rates as Predictors of Inflation: Comment

By DOUGLAS JOINES*

In a recent article in this *Review*, Eugene Fama presented evidence in support of two hypotheses: 1) that the expected real return on 1-month U.S. Treasury Bills was constant during the period January 1953 August 1971; and 2) that during this period the market for 1-month Treasury Bills efficiently predicted 1-month rates of inflation, as measured by the Consumer Price Index (*CPI*). Efficiency is taken to mean "that nominal interest rates summarize all the information about future inflation rates that is in time-series of past inflation rates" (p. 269).

In this note I raise two questions regarding Fama's results. The first is whether similar tests based on an information set somewhat broader than the past history of consumer price inflation will support Fama's findings. The second is whether given Fama's data and model his results fully support his hypotheses The answer to both questions is "no" if one accepts the accuracy of the data.

During Fama's sample period, the current monthly rate of change of consumer prices, C_t, is significantly related to the three previous monthly rates of change of wholesale prices, W_{t-1}, W_{t-2}, W_{t-3} (see the second line of Table 1).[1] As an extension of Fama's results one can ask whether in setting the nominal interest rate R_t the market utilizes all information about C_t contained in W_{t-1}, W_{t-2}, and W_{t-3}. The third line of Table 1, which reports regression coefficients and other statistics for an analogue of Fama's equation (21), sheds some light on this question. These results appear to be inconsistent with both of Fama's hypotheses. The coefficient on R_t is more than two standard errors below unity, which is inconsistent with either a constant expected real return or efficiency. Each of the coefficients on W_{t-1}, W_{t-2}, and W_{t-3} is individually different from zero at the 5 percent significance level, and a test of the hypothesis that all three coefficients are zero against the alternative that one of the three is nonzero yields an F-statistic (with 3 and 218 degrees of freedom) of 9.3 and can thus be rejected at high levels of significance These results are consistent with the hypothesis that the market ignores readily available information concerning the rate of change of consumer prices when it sets the nominal interest rate.

The nominal interest rate predicts the sum of the rate of inflation and the real return. The results reported above can be explained by a nonconstant expected real return while still maintaining the efficiency hypothesis. If the expected real return varies from month to month and is correlated with W_{t-1}, W_{t-2}, and W_{t-3}, then omitting it as an explanatory variable could result in numbers similar to those in the table, even though in setting the nominal interest rate the market takes account of all information about C_t contained in W_{t-1}, W_{t-2}, and W_{t-3}. In particular, such an omission could result in nonzero estimated coefficients for these variables even though the true coefficients are zero.

The table also shows the autocorrelation of residuals from each estimated equation at lags 1, 2, and 3, as in Fama's paper, as well as the autocorrelation at the twelfth, or seasonal, lag. The twelfth-order autocorrelation of residuals is significantly different from zero at the 5 percent level for each of the three equations. Assuming that the *CPI* and Wholesale Price Index (*WPI*) ac-

*University of Chicago I am indebted to S Ghorayeb and A. Laffer for helpful comments.

[1]No attempt is made to derive a relationship from theory Instead reference is made to an empirical relationship which existed during Fama's sample period and which the market might have taken into account in forming its anticipation of the rate of change of consumer prices.

TABLE 1 — REGRESSION RESULTS[a]

Constant	R_t	W_{t-1}	W_{t-2}	W_{t-3}	R^2	$S(e)$	$D.W.$	$\hat{\rho}_1(e)$	$\hat{\rho}_2(e)$	$\hat{\rho}_3(e)$	$\hat{\rho}_{12}(e)$
-0.00068	0.98				0.29	0.00196	1.766	0.11	0.12	-0.02	0.19
(0.00030)	(0.10)										
0.00134		0.186	0.164	0.129	0.22	0.00207	1.662	0.17	0.23	0.07	0.32
(0.00016)		(0.039)	(0.038)	(0.039)							
-0.00050	0.77	0.132	0.101	0.077	0.37	0.00186	1.998	0.00	0.08	-0.08	0.24
(0.00028)	(0.10)	(0.036)	(0.035)	(0.035)							

Notes: Numbers in parentheses are standard errors. $S(e)$ denotes standard error of residuals; $D.W.$ denotes Durbin-Watson statistic; $\hat{\rho}_j(e)$ denotes sample autocorrelation of residuals at lag j.

[a]The dependent variable in each equation is C_t, the current monthly rate of change of consumer prices.

curately measure the prices of consumer goods and wholesale commodities, this autocorrelation indicates that there is a seasonal pattern in the market's forecast errors of the rate of consumer price inflation. Such a pattern of forecast errors is strictly inconsistent with market efficiency. There are, however, at least two alternative explanations for this seasonal result as well as our earlier results. The first is that in setting the nominal interest rate the market is concerned not merely with the prices of consumer goods, and thus tries to forecast some more general index of inflation. If the seasonal pattern of the used proxy differs from the seasonal pattern of "true" inflation then there may be seasonality in the residuals.

The second explanation is that there are in fact deficiencies in the price data. Specifically, the lack of monthly sampling of all items included in the *CPI*, to which Fama refers, could produce nonzero coefficients on W_{t-1}, W_{t-2}, and W_{t-3} and nonzero value for the twelfth-order autocorrelation of the residuals. In setting the 1-month nominal interest rate, an efficient market should incorporate the effects of any true seasonal pattern in the rate of inflation. The market, however, should not react to an apparent seasonality which is merely an artifact of the method used to collect the data. Such spurious seasonality in the *CPI* should be reflected in the residuals from equations such as those reported above.[2]

[2]Question has been raised in the literature concerning the accuracy not only of the *CPI* but also of the *WPI*. Measurement errors in the *WPI* could of course produce nonzero estimated coefficients even though the true values of the coefficients are zero.

REFERENCE

E. F. Fama, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269-82.

Short-Term Interest Rates as Predictors of Inflation: On Testing the Hypothesis that the Real Rate of Interest is Constant

By CHARLES R. NELSON AND G. WILLIAM SCHWERT*

In an innovative and provocative paper in this *Review*, Eugene Fama presents evidence which appears to be consistent with the joint hypothesis that the real rate of interest, ignoring taxes, is a constant and that the market for U.S. Treasury Bills is efficient in the sense of embodying rational expectations.[1] Those findings are strikingly at variance with a long list of previous studies which seemed to support the view that the real rate of interest varies over time and can be related to economic variables such as real output, monetary policy, and so forth.[2] The methodologies which lead to these contradictory conclusions are quite different. Previous studies had followed the lead of Irving Fisher by relating nominal interest rates to distributed lags on past rates of inflation.[3] These distributed lags are generally interpreted as approximations to the market's expected rate of inflation and thus any remaining variation is attributed to variation in the real rate.[4] Fama's methodology, on the other hand, draws on the fact that the difference between the market interest rate and the *subsequently observed* rate of inflation, the *ex post* real interest rate, consists by definition of the *ex ante* real interest rate plus a pure forecasting error.[5] The hypothesis of market efficiency implies that these forecasting errors must be serially random. Thus, observing *ex post* real rates is equivalent to observing *ex ante* real rates with random measurement error. These errors of course confound the problem of identifying variation in the *ex ante* real rate

In this paper, we will argue that the relative magnitude of these measurement errors is such that the tests carried out by Fama were not powerful enough to reject the joint hypothesis that the *ex ante* real rate is a constant and expectations are rational. More powerful tests which are presented in this paper using the same data do lead to rejection of that hypothesis. Of course, rejection of the joint hypothesis could be interpreted

*Professor of economics, University of Washington, and assistant professor, Graduate School of Management, University of Rochester, respectively This research was supported in part by the National Science Foundation under grant GS 34501 and in part by a grant from the University of Chicago Graduate School of Business. We are grateful to Eugene Fama for providing us with his data on Treasury Bill yields. We also appreciated helpful criticism received from Mukhtar Ali, Truman Clark, Fama, Alan Hess, Charles Plosser, and Arnold Zellner, but we accept responsibility for all errors.

[1] See John Muth and Fama (1970) for definitions of "rationality" and "efficiency"

[2] The taxation of nominal income implies that if the before-tax real rate is constant, the after-tax real rate will vary inversely with the rate of inflation. The behavior of the after-tax rate is presumably more relevant from the viewpoint of the individual investor making consumption savings decisions

[3] Richard Roll presents a bibliography of many of these studies. Roll also argues that efficiency in markets for goods (with no costs of storing goods or shifting production through time) would imply that inflation rates be randomly distributed with mean zero, but shows that to the contrary actual inflation rates exhibit substantial autocorrelation.

[4] As John Rutledge (pp. 17–21) has pointed out, Fisher's interpretation of his distributed lags placed considerable emphasis on the indirect effects of inflation on interest rates rather than on the formation of expectations by extrapolation

[5] Fama's methodology corresponds to Reuben Kessel's methodology for studying the behavior of liquidity premiums in the term structure of interest rates In the present context, subtracting the observed inflation rate from the market interest rate removes the expected inflation rate, leaving the *ex ante* real rate and a random forecasting error. Kessel subtracted observed spot interest rates from corresponding past forward rates thereby removing the expected spot rate and leaving the liquidity premium and a random forecasting error. Under the implicit assumption of rational expectations, Kessel studied the relationship of liquidity premiums to the level of interest rates, but he did not claim to have tested that assumption.

as an indication that the *ex ante* real rate is variable or that the market is inefficient, or both. In fact, neither market efficiency nor the constancy of the real rate constitutes in itself a testable hypothesis. If we are willing to *assume* that the market is efficient, then our results suggest that variation in the real rate is of the magnitude suggested by previous studies in the Fisher tradition.[6]

I. Tests Based on Autocorrelation in the Ex Post Real Rate

Fama examines the sample autocorrelations of the *ex post* real rate as one test of his joint hypothesis. As mentioned above, the *ex post* real rate (*EPRR*) consists by definition of the *ex ante* real rate (*EARR*) plus a pure forecasting error which will be serially uncorrelated if the market is efficient. In such a market, any autocorrelation in the *EPRR* can be attributed to autocorrelation in the *EARR*. For the monthly data consisting of 1-month Treasury Bill yields used by Fama and 1-month rates of change in the Consumer Price Index (*CPI*), the sample autocorrelations for lags one through twelve computed over the period January 1953 through July 1971, are given in Table 1.[7] The autocorrelations are generally small relative to their asymptotic standard error (equal to $1/\sqrt{n}$) with the exception of the seasonal autocorrelation at lag 12 months. The Box-Pierce Q-statistic, a measure of overall autocorrelation, is fairly large and significant at the .10 level, but this is due primarily to the seasonal coefficient. Autocorrelation in the *EPRR* need not be due to market inefficiency; rather, as Fama points out, it could be due to systematic measurement errors in the Consumer Price Index or autocorrelation in the *EARR*. The fact that the sample autocorrelations of the *EPRR* are small was interpreted by Fama as an indication of market efficiency

TABLE 1—SAMPLE AUTOCORRELATIONS OF ONE-MONTH *Ex Post* REAL RATE, JANUARY 1953–JUNE 1971

Lag	Autocorrelation	Lag	Autocorrelation
1	.10	7	−.08
2	.12	8	.04
3	−.02	9	.10
4	−.01	10	.09
5	−.02	11	.03
6	−.01	12	.18

Note: Standard Error = .07; Box-Pierce $Q = 19.4$ (χ^2 with 12 degrees of freedom).

and constancy of the *EARR*. However, lack of serial correlation in the *EPRR* is also consistent with variation in the *EARR* which is purely random in nature and also with market inefficiency in the form of forecast errors which are larger than necessary given available information. What we wish to demonstrate here is that even the rather low autocorrelation observed in the *EPRR* could in fact be indicative of rather strong autocorrelation and sizable variation in the *EARR* since the latter is being overlaid with forecast errors when we observe the former.

Suppose that the *EARR* rather than being constant is a stochastic process with first-order serial correlation coefficient ϕ. Denoting the market's forecasting error for the rate of inflation by ϵ_t, the *EARR* by i_t, and the EPRR by r_t, we have

$$r_t \equiv i_t - \epsilon_t \tag{1}$$

It is easy to show that in an efficient market the first-order serial correlation coefficient for the *EPRR* will be related to ϕ by

$$\text{Corr}(r_t, r_{t\pm 1}) = \frac{\phi\sigma_i^2}{\sigma_i^2 + \sigma_\epsilon^2} \tag{2}$$

where σ_i^2 and σ_ϵ^2 denote the variances of the *EARR* and the forecast error, respectively. It should be clear that first-order autocorrelation in the *EPRR* may be considerably less than ϕ if the variance of forecast errors is large relative to the variance of the *EARR*.

It is interesting to consider what combinations of autocorrelation and variance for the *EARR* are consistent with low auto-

[6]For example, see the calculated real rate series by William Yohe and Denis Karnosky.

[7]We use the more familar rate of inflation, the natural logarithm of the price relative, instead of the discrete percent change in purchasing power used by Fama. Also, our estimated autocorrelations are computed as correlation coefficients whereas Fama computed regression coefficients.

TABLE 2—VALUES OF ϕ AND σ_w^2 AND CORRESPONDING VARIANCES AND STANDARD DEVIATIONS FOR THE *EARR* AS AN ANNUAL PERCENTAGE RATE CONSISTENT WITH $\text{Corr}(r_t, r_{t+1}) = .10$ and $\sigma_\epsilon^2 = 4.32$

ϕ	Variance of *EARR*, σ_i^2	Standard Deviation of *EARR*, σ_i	σ_w^2
.4	1.44	1.20	1.21
.5	1.08	1.04	.806
.6	.864	.924	.547
.7	.720	.852	.374
.8	.619	.780	.230
.9	.547	.732	.102
.95	.508	.708	.049
.99	.485	.696	.010
.999	.481	.696	.001

Note: The statistic σ_w^2 is computed under the assumption $i_t = \phi i_{t-1} + w_t$.

correlation in the *EPRR*, say the .10 observed during the sample period. To do this we require an estimate of σ_ϵ^2. An upper bound estimate of σ_ϵ^2 would be the variance of the *EPRR*, 5.18 on an annualized percentage basis, since that estimate would attribute all variation in the *EPRR* to the forecast error. Estimates derived in Section III under less restrictive assumptions suggest that a conservative estimate (in the sense of yielding smaller values of ϕ and σ_i^2) for purposes of illustration would be 4.32 percent on an annual basis. Corresponding values of ϕ and of the variance and standard deviation of the *EARR* in annualized percentage terms are given in Table 2. It is apparent that low autocorrelation in the *EPRR* may be compatible with substantial variation and autocorrelation in the *EARR*. For example, values of ϕ of .99 or .999 which would imply behavior approaching nonstationarity are also consistent with a standard deviation in the *EARR* of about .7 percent or a .95 probability interval of nearly 3 percent, assuming normality.

Even if the *EARR* were a random walk having no long-run mean and unbounded variance, we might very well observe sample autocorrelations as low as those in Table 1. Denoting the hypothetical random steps in the *EARR* by w_t, so that $i_t = i_{t-1} + w_t$, it is easy to show that a random walk in the *EARR* would imply that the *EPRR* is generated by the process

$$r_t = r_{t-1} + w_t - \epsilon_t + \epsilon_{t-1} \tag{3}$$

which could be thought of as a random walk with an autocorrelated disturbance. The *theoretical* autocorrelations for r are undefined (as they are for i) but *sample* autocorrelations are of course readily computed for any finite data series and their magnitude will depend on the autocorrelation properties of the moving average ($w_t - \epsilon_t + \epsilon_{t-1}$). The analysis given by George Box and Gwilym Jenkins implies that the sample autocorrelations of r_t will tend to be quite small if the variance of w_t is small relative to the variance of the forecasting errors ϵ_t.[8] In particular, a value of .011 for the variance of w_t would be implied by our estimate of σ_ϵ^2 and the observed sample first-order autocorrelation in r. To interpret this, note the last column of Table 2 which gives the variance of w_t, denoted σ_w^2, in the stationary process $i_t = \phi i_{t-1} + w_t$, $|\phi| < 1$, for the accompanying values of ϕ and σ_i^2. The value of σ_w^2 computed from the formula given by Box and Jenkins under the random walk assumption is in fact larger than the values in Table 2 for the cases $\phi = .99$ and .999.

We conclude in this section that tests of Fama's hypothesis based on sample autocorrelations of the *EPRR* will have little power against alternatives which specify economically plausible variation and autocorrelation in the *EARR* or even the alternative hypothesis of a random walk in the *EARR*.

II. Regression Tests

Fama also presented tests of the joint hypothesis of a constant real rate and an efficient market based on regressions of realized inflation rates ρ_t, on the market interest rate R_t, and past rates of inflation. If $\hat{\rho}_t$ represents a piece of information about ρ_t which is available to the market at the beginning of period t, for example, a past

[8] Box and Jenkins (pp. 200–01) evaluate the ratio of the expectation of the sample autocovariance at lag one to the expectation of the sample variance, both conditional on the initial observation in the sample, for a process similar to (3).

rate of inflation, then the regression of ρ_t on R_t and $\hat{\rho}_t$ would yield a nonzero coefficient for $\hat{\rho}_t$ only if the market were either inefficient in its use of available information or if the predictive ability of R_t were distorted by underlying variation in the *EARR*. Fama chose ρ_{t-1} as a particular $\hat{\rho}_t$ and found that the coefficient of ρ_{t-1} was small and not significant. The power of such a test will be low, however, if the $\hat{\rho}_t$ chosen contains little information about ρ_t.

In order to provide a more powerful test of the joint hypothesis of market efficiency and the constancy of the *EARR*, we use the past rates of inflation to construct a $\hat{\rho}_t$ which is an optimal extrapolative predictor using the methodology of Box and Jenkins. While this approach will not generally yield a fully rational forecast of ρ_t, it will generally be a more efficient predictor than ρ_{t-1}.[9]

The weights assigned to R_t and $\hat{\rho}_t$ in a composite prediction of ρ_t can be related to measures of market efficiency and variation in the *EARR*. Define u_t as the component of information embodied in the market forecast but not in the extrapolative predictor, and v_t as the component of information embodied in the extrapolative predictor but ignored by the market. These random variables have variances σ_u^2 and σ_v^2, respectively, and are uncorrelated with each other. The probability limits of the weights assigned to R_t and $\hat{\rho}_t$ in the composite predictor of ρ_t are approximately $\sigma_u^2/(\sigma_u^2 + \sigma_v^2 + \sigma_r^2)$ and $(\sigma_v^2 + \sigma_r^2)/(\sigma_u^2 + \sigma_v^2 + \sigma_r^2)$, respectively, if the *EARR* is uncorrelated with the anticipated rate of inflation. Thus, the relative weight assigned to $\hat{\rho}_t$ will be greater than zero if variation in the *EARR* (σ_r^2) is nonzero or if the market ignores any information available from past inflation rates (measured by σ_v^2).[10]

The autocorrelation structure of the *CPI* monthly inflation series for January 1953 through July 1971 suggests that the series may reasonably be represented as a first-order moving average process in its first differences, implying nonstationary behavior in the rate of inflation. The estimated model for the 2/53–7/71 period is

$$(4)\qquad (1 - B)\rho_t = \underset{(.0179)}{.0222} + (1 - \underset{(.029)}{.894}B)e_t$$

$$\hat{\sigma}_e = 2.408$$

where e_t is a sequence of residuals which are the one-step-ahead forecast errors for this model and B is the lag operator. The forecast $\hat{\rho}_t$ of ρ_t implied by this model may be written apart from a constant as

$$(5)\qquad \hat{\rho}_t = .11\rho_{t-1} + (.89)(.11)\rho_{t-2} + \dots + (.89)^j(.11)\rho_{t-j-1} + \dots = \sum_{j=0}^{\infty} \theta^j(1 - \theta)\rho_{t-j-1}$$

where $\theta = .89$. Note that the weight given to ρ_{t-1} in this forecast is very small; interestingly, it is the same as the regression coefficient of ρ_{t-1} in Fama's regressions of ρ_t on R_t and ρ_{t-1} for monthly data.[11] Also note that the sample standard deviation of the forecast errors e_t is larger than our estimate of the standard deviation of market forecast errors, the residuals from the regression of ρ_t on R_t in Table 3, which is consistent with the hypothesis that the market utilizes information beyond that available from past inflation rates alone. Reestimation of (4) over the subperiods 1/53–2/59, 3/59–7/64, and 8/64–7/71 revealed little variation in the estimates of the moving average parameter θ. This model can be used to predict either the level ρ_t, or the change $\rho_t - \rho_{t-1}$ in the rate of inflation; in either case, the one-step-ahead prediction errors are the disturbances e_t.

[9] Rutledge (pp. 57–61) has carried this reasoning one step further in the context of a model in which expectations of inflation depend on past money growth rates as well as past inflation rates. Regressions of the *EPRR* associated with forward Treasury Bill yields on these variables led Rutledge to conclude that the joint hypothesis of rational expectations and constancy of the (forward) real rate could not be rejected.

[10] It should be noted that serially correlated measurement errors in the *CPI* might be predicted by $\hat{\rho}_t$ but ignored by R_t, since an efficient market should ignore these measurement errors in setting interest rates. This possibility is observationally equivalent to the possibility that the market ignores information contained in past inflation rates.

[11] See Fama (1975), Table 4, p. 276, where the regression coefficient for ρ_{t-1} is given as .11 for the 1/53–7/71 sample period.

TABLE 3

α	β	γ	$\hat{\sigma}_e$	R^2	D.W.
A. Composite Predictors of the Rate of Inflation: 2/53–7/71					
−.775	.969		2.347	.292	1.81
(.358)	(.102)				
−.641	.651	.383	2.322	.310	1.93
(.359)	(.165)	(.158)			
B. Composite Predictors of the Change in the Rate of Inflation: 2/53–7/71					
−.774	.889		2.333	.458	2.07
(.167)	(.065)				
−.546	.633	.317	2.320	.466	2.03
(.206)	(.152)	(.170)			

Note: The prediction equations are:
A. $\rho_t = \alpha + \beta R_t + \gamma\hat{\rho}_t + e_t$
B. $(\rho_t - \rho_{t-1}) = \alpha + \beta(R_t - \rho_{t-1}) + \gamma(\hat{\rho}_t - \rho_{t-1}) + e_t$

(Standard errors in parentheses.)

Regressions of ρ_t on R_t and $\hat{\rho}_t$ for the 2/53 to 7/71 period in Table 3 indicate that the extrapolative predictor has a large and significant weight in the composite predictor for ρ_t. Since there is evidence that ρ_t is not stationary, we also estimate regressions which are designed to predict the change in the rate of inflation $\rho_t - \rho_{t-1}$, using predictors of the change: $R_t - \rho_{t-1}$ and $\hat{\rho}_t - \rho_{t-1}$. Again, the coefficient of the extrapolative predictor is large though not as strongly significant as for the level of the rate of inflation.

These composite prediction tests are subject to criticism on the grounds that the extrapolative predictors are computed using time-series models estimated over the entire sample period and thus possibly use information not fully available to the market. This is not likely to be important since estimates of the parameters of the time-series models are quite stable when estimated over subperiods. As a check on the sensitivity of the results we computed $\hat{\rho}$ sequentially, using data for 1/43–12/52 to calculate forecasts for 1953, then data for 1/44–12/53 to calculate forecasts for 1954, and so forth.[12] We found that the differences in the predictive regressions were too small to alter our basic conclusions. Ideally, the composite weights should be estimated simultaneously with the parameters of the time-series model by specifying a dynamic regression (transfer function) model which includes pure interest rate and pure time-series prediction as special cases. The resulting point estimates and their standard errors would reflect appropriately the information contained in the data. Taking this approach we embedded the interest rate in the time-series model for inflation and obtained the following results for the 2/53–7/71 period:

$$(6)\quad (1 - B)\rho_t = \underset{(.0211)}{.0199} + \underset{(.262)}{.577}(1 - B)R_t + (1 - \underset{(.032)}{.876}B)e_t$$

$$\hat{\sigma}_e = 2.430$$

If the coefficient of $(1 - B)R_t$ is zero, the model reduces to the pure extrapolative model; on the other hand, if the coefficient of e_{t-1} is unity, the model reduces the regression of ρ_t on R_t. Both of these polar cases are rejected by the data.[13]

These composite prediction regressions strongly suggest that the *EARR* is not constant since, as we have seen, in an efficient market "naive" predictors such as $\hat{\rho}_t$ should add nothing to the predictive power of the market interest rate. These results are consistent, however, with the hypothesis that part of the variation in R is due to the nonpredictive variation of the *EARR*. Our results also suggest that the market draws on information beyond that available from past inflation rates ($\sigma_u^2 > 0$) since the weight given to the interest rate is large and significant.

Similar results are obtained when the time-series model is expanded to account for seasonality which is present in the *CPI*.[14] Since this seasonal autocorrelation is

[12]This is essentially the strategy followed by Patrick Hess and James Bicksler in their analysis of Fama's regression tests.

[13]The observant reader may have noted that $\hat{\sigma}_e$ is larger in (6) where R_t is included than in the pure time-series model (4) where R_t does not appear. This is due to a difference in procedures for computing residuals in non-linear least squares which involved "backforecasting" presample data in the case of (4) but not in (6).

[14]The seasonal model is a multiplicative seasonal moving average model with seasonal and ordinary dif-

(ploited by the model to increase the pre-ictive power of $\hat{\rho}$, it is not surprising that ie weight given $\hat{\rho}$ in composite predictions icreases from the values of .383 and .317 ı Table 3 to .448 and .627, respectively, 'ith roughly corresponding reductions in ie weights given to R. When R is em-edded in the seasonal time-series model, it ontinues to exert a significant influence. ome of the seasonality in the *CPI* may be ue to the fact that prices of certain items re not sampled every month. For this rea-on we feel it is more conservative to em-hasize the results which do not depend on redicting seasonal variation.

III. Identification of Estimates of the Variance of the Ex Ante Real Rate of Interest

Since the joint hypothesis of market effi-:iency and constancy of the *EARR* is re-ected by the composite prediction tests, it s a matter of some interest to identify an :stimate of the variance of the *EARR* in .erms of observable variables. We approach :his problem by examining the relationships inking the variances and covariances of ob-served variables to those of the unobserved components of those variables including the *EARR* We begin with the definitions

$$R_t = \rho_t^* + i_t \qquad (7)$$
$$\rho_t = \rho_t^* + \epsilon_t$$

where R_t is the observed nominal interest rate, ρ_t^* is the unobserved market forecast of the inflation rate ρ_t, i_t is the unobserved *EARR* as before, and ϵ_t is the market fore-cast error as before. The variances of R and ρ and their covariances can be estimated from data. From these we would like to solve for the variances of i, ϵ, ρ^*, and their covariances. For a solution to be possible, we need to impose enough restrictions to re-duce the number of unknowns to three. Under the *assumption* that the market is effi-cient, we can eliminate the covariances be-tween i and ϵ and between ρ^* and ϵ, since both i and ρ^* represent *ex ante* information. If we are willing to assume further that the covariance between ρ^* and i is zero (no "Mundell Effect"), then the covariance be-tween R and ρ is simply the variance of ρ^* and thus

$$\hat{\sigma}_i^2 = \hat{\sigma}_R^2 - \hat{\sigma}_{R\rho} \qquad (8)$$
$$\hat{\sigma}_\epsilon^2 = \hat{\sigma}_\rho^2 - \hat{\sigma}_{R\rho}$$

The same relations hold with regard to vari-ances and covariances of changes in the rate of inflation $(\rho_t - \rho_{t-1})$ and predicted changes in the rate of inflation $(R_t - \rho_{t-1})$ and $(\rho_t^* - \rho_{t-1})$. Since there is some doubt about the stationarity of ρ_t and R_t and therefore about the existence of variances and covariances, computations in terms of changes in the rate of inflation and $(R_t - \rho_{t-1})$ may be preferable. Sample moments for these variables from the data used by Fama imply

$$\hat{\sigma}_i^2 = .642 \qquad (9)$$
$$\hat{\sigma}_\epsilon^2 = 4.85$$

It is interesting to note that this estimate of σ_i^2, the variance of the *EARR*, is quite close to those presented in Table 2 which are associated with a fairly strongly autocorre-lated *EARR*.

In a well-known paper, Robert Mundell has argued that an increase in the antici-pated rate of inflation will be accompanied by a fall in the *ex ante* real rate. A full interpretation of Mundell's result in a dynamic context goes beyond the scope of this paper, but it does suggest that the co-variance over time between the expected rate of inflation or expected changes in the rate of inflation and the *ex ante* real rate may be nonzero. It is clear from the above analysis that additional information must be added if an estimate of this covariance is to be identified. Additional information is available from the variance of $(\hat{\rho}_t - \rho_{t-1})$ and its covariance with $(R_t - \rho_{t-1})$ and $(\rho_t - \rho_{t-1})$, although the covariance of $(\hat{\rho}_t - \rho_{t-1})$ with i, $(\rho_t^* - \rho_{t-1})$ and ϵ will be additional unknowns. Market efficiency im-plies that the third of these covariances is zero, and if we are willing to assume that

ferencing The moving average parameters are both about .9, so that the seasonal and nonseasonal parts of the model are very similar. The parameter estimates were again very stable over subperiods. We benefited from discussions with Charles Plosser on seasonality in the *CPI*.

the covariance between $(\hat{\rho}_t - \rho_{t-1})$ and i is the same as that between $(\rho_t^* - \rho_{t-1})$ and i, then it is easy to show that solution is feasible. The resulting estimates are

$$(10) \quad \hat{\sigma}_{(\rho_t^* - \rho_{t-1}),i} = \hat{\sigma}_{(\hat{\rho}_t - \rho_{t-1}),i} = .058$$
$$\hat{\sigma}_i^2 = .585$$
$$\hat{\sigma}_\epsilon^2 = 4.91$$

The estimated covariance between expected inflation and the *EARR* is positive and so small that it causes only a slight change in $\hat{\sigma}_i^2$ and $\hat{\sigma}_\epsilon^2$ relative to equations (9).

This analysis of the variances and covariances of the unobservable components of the interest rate and inflation can be extended to the interpretation of regressions of ρ_t on R_t as reported by Fama. As Fama noted, a slope coefficient of unity in these regressions would be consistent with the hypothesis that the *EARR* is constant. Since the estimated coefficient in this regression (Table 3, Panel A) is .969 with a standard error of .102, the hypothesis seems to be supported by the monthly data. We have argued previously in this paper that the rate of inflation is nonstationary; thus for purposes of discussing probability limits it is more appropriate to work with changes in the rate of inflation and predicted changes. The slope in the change regression (Table 3, Panel B) is .889 with a standard error of .065, so this alternative regression would seem to cast doubt on the constant *EARR* hypothesis. The probability limit of this slope is the ratio of the covariance between $(\rho_t - \rho_{t-1})$ and $(R_t - \rho_{t-1})$ over the variance of $(R_t - \rho_{t-1})$ which is equivalent to

$$(11) \quad plim\, \hat{\beta}_{(\rho_t - \rho_{t-1}),(R_t - \rho_{t-1})} = \frac{\sigma^2_{(\rho_t^* - \rho_{t-1})} + \sigma_{(\rho_t^* - \rho_{t-1}),i}}{\sigma^2_{(\rho_t^* - \rho_{t-1})} + \sigma_i^2 + 2\sigma_{(\rho_t^* - \rho_{t-1}),i}}$$

This expression will deviate from unity if variation in the *EARR* is nonzero. The estimated slope of .889 can be interpreted, then, as being consistent with the small positive estimate of $\sigma_{(\rho_t^* - \rho_{t-1}),i}$ given in (10), and variation in the *EARR* which is small *relative* to the variation in the predicted change in the rate of inflation.

These results change in an interesting way if the $\hat{\rho}$ used in the calculations is generated by the seasonal time-series model. In this case we find

$$(12) \quad \hat{\sigma}_{(\rho_t^* - \rho_{t-1}),i} = \hat{\sigma}_{(\hat{\rho}_t - \rho_{t-1}),i} = -1.03$$
$$\hat{\sigma}_i^2 = 1.67$$
$$\hat{\sigma}_\epsilon^2 = 3.82$$

Thus there does seem to be a Mundell Effect in seasonal variation.[15] This negative covariance between the *EARR* and the expected change in the rate of inflation is sufficient to increase substantially the estimate of the variance of the *EARR* implying a .95 probability range of about 5 percent on an annual basis.

All of these estimates of the variance of the *ex ante* real rate of interest are consistent with the results of the previous sections of the paper. Under a variety of assumptions about the relationships among the unobservable components of the market interest rate and the rate of inflation, we estimate the standard deviation of the *ex ante* monthly real rate to be between .7 and 1.3 expressed as an annualized percentage rate.

IV. Conclusion

As Fama noted, it is not possible to test the efficiency of the market for U.S. Treasury Bills without some model for the behavior of the *ex ante* real rate of interest. Conditional on the assumption that the *EARR* is constant, Fama used the sample autocorrelation function of the *ex post* real rate and the regression relationship of rates of inflation on market interest rates and the prior rate of inflation to test market efficiency. However, those tests may not have been powerful enough to lead to a rejection of the joint hypothesis of market efficiency and constancy of the *EARR* even if the *EARR* varies as much as has been indicated by previous studies.

We have shown that the autocorrelation

[15] Hess and Bicksler test for and find a negative relationship between the level of inflation and the *EARR*. We interpret Mundell's analysis as one which applies to short-run shifts in expectations rather than long-run movements in the level of expectations such as have occurred over the postwar period.

function of the *ex post* real rate of interest may be quite close to zero at all lags, even if the *ex ante* real rate varies substantially and is highly autocorrelated, because of the relatively large variance of errors in expectations of inflation. In fact, the autocorrelations of the *EPRR* may be small even if the *EARR* is a nonstationary stochastic process.

By examining the time-series properties of rates of inflation computed from the *CPI*, we have shown that *individual* past inflation rates contain very little information about future rates of inflation, suggesting that Fama's test based on the regression of the inflation rate on the market interest rate and the prior inflation rate is not a powerful test. To increase the power of the regression test we use the time-series properties of the rate of inflation to construct an optimal predictor of inflation based on the past history of inflation rates. The coefficient of the time-series predictor is large and significant in a composite prediction regression equation which also includes the market interest rate. This result would only occur if the market were inefficient in assimilating information contained in past inflation rates, or if variation in the *EARR* distorts the predictive variation in the market interest rate, or both. Thus, by making more efficient use of the information about future inflation which is contained in past inflation rates, we are able to reject Fama's joint hypothesis.

In order to get estimates of the magnitude of the variability of the *ex ante* real rate, we *assume* market efficiency and use the relationships among the unobservable components of the market interest rate and the observed inflation rate to identify estimates of the variance of the *EARR*. These estimates are comparable to those which have been derived by others working in the Fisher tradition and they indicate substantial variability in the monthly *EARR*.

There are many important issues related to measurement errors in the *CPI*. Certain kinds of errors can yield implications which are observationally equivalent to market inefficiency or variation in the *EARR*. Other types of measurement errors could have opposite effects on tests of Fama's hypothesis; for example, since the *CPI* is collected in mid-month the beginning-of-month interest rate has a two week lead on the inflation rate computed from the *CPI*. However, a full treatment of the effects of measurement error in the *CPI* is beyond the scope of this paper. In all of our analysis we have taken the data at face value and focused on the statistical issues raised by Fama's tests.[16]

Finally, while we are obliged to reject Fama's joint hypothesis, we do feel that Fama made an important methodological contribution to the literature on interest rates and inflation by shifting attention from regressions of interest rates on past inflation rates to relationships between interest rates and subsequently observed inflation rates. Our analysis does suggest that expectations of inflation have accounted for most of the variation in short-term interest rates during the postwar period, and that those expectations embody significant information beyond that contained in past inflation rates alone.

[16]However, we do not agree with Fama (1975, p. 274) that the adequacy of the data should be judged, *ex post*, by the outcome of tests of his hypothesis.

REFERENCES

George E. P. Box and Gwilym M. Jenkins, *Time Series Analysis*, San Francisco 1970.

______ **and D. Pierce,** "Distribution of Residual Autocorrelations in Integrated Moving Average Time Series Models," *J. Amer. Statist. Assn.*, Dec. 1970, *65*, 1509-26.

E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, *35*, 383-417.

______, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269-82.

Irving Fisher, *The Theory of Interest*, New York 1930.

P. Hess and J. Bicksler, "Capital Asset Prices versus Time Series Models as Predictors of Inflation," *J. Finance. Econ.*, Dec. 1975, *2*, 341-60.

Reuben A. Kessel, *The Cyclical Behavior of the Term Structure of Interest Rates*, New York 1965.

R. Mundell, "Inflation and Real Interest," *J. Polit. Econ.*, June 1963, *71*, 280–83.

J. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

R. Roll, "Interest Rates on Monetary Assets and Commodity Price Index Changes," *J. Finance*, May 1972, *27*, 251–77.

John Rutledge, *A Monetarist Model of Inflationary Expectations*, Lexington 1974.

W. P. Yohe and D. S. Karnosky, "Interest Rates and Price Level Changes, 1952–69," *Fed. Reserve Bank St. Louis Rev.*, Dec. 1969, *51*, 18–38.

Interest Rates and Inflation: The Message in the Entrails

By EUGENE F. FAMA*

In "Short-Term Interest Rates as Predictors of Inflation," I tested the joint hypotheses that (i) expected real returns on U.S. Treasury Bills are constant through time, and (ii) the Treasury Bill market is efficient in the limited sense that interest rates on bills are based on assessments of expected future inflation rates that properly use the information in past inflation rates. I conclude that these hypotheses are consistent with the data for the 1953–71 period.

John Carlson, Douglas Joines, and Charles Nelson and G. William Schwert present tests that reject my model at standard levels of statistical significance. Similar negative results were reported in an earlier comment by Patrick Hess and James Bicksler.

Taking the data at face value, the evidence that one can devise tests that reject my model is incontestable. However, many of the apparent shortcomings of the model could be manifestations of systematic measurement errors in the data. Moreover, the deviations from the model, uncovered after thorough searching of the data, seem small enough that the model remains a useful approximation to the world.

I. The Model and Tests: A Brief Review

Let R_t be the nominal interest rate quoted at the end of month $t-1$ on a Treasury Bill that matures at the end of month t. Let $E(\tilde{r})$ be the expected value of the real return on the bill for the month, with $E(\tilde{r})$ assumed to be constant through time. The real return $\tilde{r}_t$,

$$(1) \qquad \tilde{r}_t = R_t - \tilde{\Delta}_t$$

is a random variable, indicated by the tilde, because the inflation rate $\tilde{\Delta}_t$ for month t is uncertain at the end of month $t-1$.[1]

With this notation, my model can be expressed as

$$(2) \qquad R_t = E(\tilde{r}) + E(\tilde{\Delta}_t \mid \Delta_{t-1}, \Delta_{t-2}, \ldots)$$

that is, the nominal interest rate is the constant expected real return plus the expected inflation rate implied by the information in past inflation rates. Since $E(\tilde{r})$ is the same for all t, all variation through time in the nominal interest rate directly reflects variation in the expected inflation rate. Alternatively, rearranging (2) we have

$$(3) \qquad E(\tilde{\Delta}_t \mid \Delta_{t-1}, \Delta_{t-2}, \ldots) = -E(\tilde{r}) + R_t$$

This way of expressing the model emphasizes its implication that the interest rate R_t set at $t-1$ summarizes all the information about the inflation rate from $t-1$ to t which is in the time-series of past inflation rates. Putting this statement in the form most relevant for testing, we have

$$(4) \qquad E(\tilde{\Delta}_t \mid R_t, \Delta_{t-1}, \Delta_{t-2}, \ldots) = -E(\tilde{r}) + R_t$$

The conditional expected value $E(\tilde{\Delta}_t \mid R_t, \Delta_{t-1}, \Delta_{t-2}, \ldots)$ is the regression function of $\tilde{\Delta}_t$ on R_t and on past inflation rates. Although these past inflation rates may contain information relevant for assessing the expected inflation rate from $t-1$ to t, equation (4) says that all of this information is summarized in the nominal rate of interest R_t. Once R_t is set at $t-1$, the information in past inflation rates becomes redundant for assessing the expected value of $\tilde{\Delta}_t$.

*Professor of finance, Graduate School of Business, University of Chicago. The financial support of the National Science Foundation and the comments of N. Gonedes, M. Miller, M. Scholes, G. W. Schwert, A. Zellner, and especially H. Roberts are gratefully acknowledged.

[1]When continuous rates of change are used, the inflation rate is the negative of the rate of change in purchasing power, the measure of price change used in my 1975 paper and in Carlson's comment. In line with Nelson-Schwert and Joines, the more familiar inflation rate is used here.

Since (4) is a proposition about the form of a regression function, it can be tested with regression techniques. In any estimated regression of $\tilde{\Delta}_t$ on R_t and any function of past inflation rates, the coefficient of R_t should be statistically indistinguishable from 1.0, the coefficients on variables representing historical inflation rates should be indistinguishable from zero, and the residuals from the regression should be serially uncorrelated for all lags. My main tests of these propositions come from estimates of

$$\tilde{\Delta}_t = \alpha_0 + \alpha_1 R_t + \tilde{\epsilon}_t \tag{5}$$

With nominal interest rates on 1-month Treasury Bills used for R_t and with $\tilde{\Delta}_t$ estimated from the U.S. Consumer Price Index (CPI), the estimate of α_1 from monthly data for January 1953 to July 1971 is .98, with a standard error of .10, and the serial correlations of the residuals from the regression are generally within standard sampling limits.

Because $\tilde{\Delta}_t$ and R_t seem to follow processes that are nonstationary in their means, Nelson and Schwert are suspicious of the regression of $\tilde{\Delta}_t$ on R_t. Like any regression function, however, the role of (4), estimated from (5), is to transform $\tilde{\Delta}_t$ into a time-series of random disturbances $\tilde{\epsilon}_t$. Most important, the disturbances should be serially uncorrelated. Since the residuals from the estimated regression seem well behaved, the regression is well specified and the manipulations of the data undertaken by Nelson and Schwert to achieve stationarity are unnecessary.

Carlson is also unnecessarily worried by the nonstationarity of $\tilde{\Delta}_t$ and R_t. One of his main criticisms is that the regression of the inflation rate on the interest rate works well because of common trends, that is, mean nonstationarities, in the two variables. My model implies, however, that any trends in the expected inflation rate should show up in the interest rate. If the two variables did not follow common trends, the model would be rejected.

I also estimate the autocorrelations of the real return $\tilde{r}_t = R_t - \tilde{\Delta}_t$. If the expected real return is constant through time, and if the market's forecast of $\tilde{\Delta}_t$ correctly incorporates the information in past inflation rates, then the market will set R_t so as to offset any autocorrelation in $\tilde{\Delta}_t$ with the result that the real return $\tilde{r}_t$ is serially uncorrelated. The first twelve estimated autocorrelations of monthly values of $\tilde{\Delta}_t$ are on the order of .3 (which indicates that past inflation rates carry nontrivial information about future inflation rates), while the autocorrelations of the real returns are close to zero (which is consistent with the proposition that the nominal rate R_t set at time $t-1$ appropriately captures the information about Δ_t which is in past inflation rates)

II. The Challenge from Optimal Time-Series Predictors

Given a constant expected real return, the autocorrelation function of the residuals from the estimates of (5) and the autocorrelation function of real returns contain evidence relevant for determining whether the market appropriately uses the information in past inflation rates when it assesses expected future inflation rates. Simply looking at autocorrelations individually, however, as I did, may not effectively isolate the information they contain.

To get a better measure of the information in past inflation rates, Nelson and Schwert, like Hess and Bicksler, use the techniques of George Box and Gwilym Jenkins to estimate an optimal time-series forecaster of $\tilde{\Delta}_t$, call it $\hat{\Delta}_t$, from the past values $\Delta_{t-1}, \Delta_{t-2}, \ldots$. They then find that an estimated regression of $\tilde{\Delta}_t$ on both R_t and the optimal time-series predictor $\hat{\Delta}_t$ produces a coefficient for $\hat{\Delta}_t$ which is more than two standard errors from zero. Apparently $\hat{\Delta}_t$ contains information about $\tilde{\Delta}_t$ that is not included in the interest rate R_t set at time $t-1$. However, the coefficient of determination in the multiple regression is .31, as compared to .29 for the regression that includes R_t alone. Including $\hat{\Delta}_t$ reduces the standard error of the regression residuals from 2.347 to 2.322.

The Box-Jenkins forecaster $\hat{\Delta}_t$ is the end result of a thorough search through the data

whose object is to capture most effectively the information in past inflation rates. Nevertheless, this time-series forecaster makes a small contribution to the prediction of the inflation rate beyond that given by the interest rate alone. The simple model which postulates an efficient market and a constant expected real return seems to remain a useful approximation to the world.

III. The Information in the WPI

In a regression of the inflation rate (estimated from the *CPI*) on the interest rate and three lagged inflation rates (estimated from the U.S. Wholesale Price Index (*WPI*)), Joines estimates coefficients for the lagged inflation rates that are more than two standard errors from zero. In replicating and extending his results, I find that the lags of *WPI* are proxying for lagged inflation effects similar to those uncovered by Nelson-Schwert and Hess-Bicksler. The evidence is in Tables 1 and 2. Table 1 shows means, standard deviations, and autocorrelations for different variables that appear in my work and in the comments under discussion. Table 2 summarizes estimated regressions of the *CPI* inflation rate Δ_t on various combinations of lagged values of Δ_t, the interest rate R_t, and lags of the *WPI* inflation rate, indicated as $W_{t-\tau}$.

The first four regressions in Table 2A compare models for Δ_t based only on lagged values of Δ_t with models of Δ_t based only on lagged values of W_t. It is evident that lags of the *CPI* inflation rate give a better model for the *CPI* inflation rate than lags of the *WPI* inflation rate. When Δ_t is regressed on Δ_{t-1}, Δ_{t-2}, and Δ_{t-3}, the residual autocorrelations (the ρ_τ) are generally within two standard errors of zero. Adding a twelfth-order lag or seasonal term Δ_{t-12} further improves the model for Δ_t. The coefficient of determination (R^2) goes from .203 to .251, and all of the residual autocorrelations are within two standard errors of zero. In constrast, the regressions of Δ_t on only lags of the *WPI* inflation rate have unsatisfactory diagnostics. The residual autocorrelations are large and they

TABLE 1 MEANS, STANDARD DEVIATIONS, AND FIRST TWELVE ESTIMATED AUTOCORRELATIONS FOR VARIOUS VARIABLES

	Variable						
	Δ_t (*CPI*)	r_t (real return)	W_t (*WPI*)	R_t	$R_t - R_{t-1}$	$(E/P)_{t-6}$	$(E/P)_{t-6} - (E/P)_{t-6-1}$
$\bar{X}$	.00188	.00074	00123	.00262	.00001	.55	.00001
$s(\bar{X})$	.00233	.00196	.00355	00129	.00032	.00836	.00248
ρ_1	.37	.11	.03	.97	−.25	.96	−.24
ρ_2	.36	.12	.13	.95	.06	.93	.07
ρ_3	.27	−.02	−.07	.93	.01	.90	.04
ρ_4	.30	−.01	.13	.91	.15	.87	.12
ρ_5	.28	−.03	−.02	.88	−.03	.82	03
ρ_6	.28	−.02	.19	.85	−.05	.77	−.03
ρ_7	.25	−.07	.07	.83	−.12	.73	.15
ρ_8	.33	.05	.17	.81	.09	.67	−.09
ρ_9	.35	.10	−.02	.78	.06	.61	.07
ρ_{10}	.33	.10	10	.74	−.22	56	−.07
ρ_{11}	.26	.03	.04	.72	−.05	.50	.06
ρ_{12}	.36	.19	.16	.70	.08	.45	−.13
$s(\rho)$	.07	.07	.07	.07	.07	.07	.07
BP	262	21	32	1,898	39	1,502	31

Note: The ρ_τ are Box-Jenkins estimates of autocorrelations; $s(\rho)$ is the approximate standard error of each ρ_τ under the hypothesis that the true autocorrelation is zero; and BP is the Box-Pierce statistic for the first twelve estimated autocorrelations. Under the hypothesis that the true autocorrelations are zero, the expected value of BP and its standard error are 12 and 5.5. These are the relevant mean and standard error of BP throughout the tables of this study.

TABLE 2A—ESTIMATED REGRESSIONS OF THE *CPI* INFLATION RATE (Δ_t) ON LAGS OF THE *CPI* INFLATION RATE ($\Delta_{t-\tau}$), ON LAGS OF THE *WPI* INFLATION RATE ($W_{t-\tau}$), AND ON THE ONE-MONTH INTEREST RATE (R_t) SET AT THE END OF MONTH $t-1$

								R^2	$s(\epsilon)$
1.	.00082	$+.239\Delta_{t-1}$	$+.261\Delta_{t-2}$	$+.088\Delta_{t-3}$				.203	.00203
	(.00020)	(.067)	(.064)	(.066)					
2.	.00066	$+.198\Delta_{t-1}$	$+.215\Delta_{t-2}$	$+.051\Delta_{t-3}$	$+.231\Delta_{t-12}$			.251	.00197
	(.00021)	(.067)	(.067)	(.068)	(.065)				
3.	.00137	$+.189W_{t-1}$	$+.144W_{t-2}$	$+ 126W_{t-3}$				.187	.00205
	(.00016)	(.039)	(.039)	(.039)					
4.	.00132	$+ 182W_{t-1}$	$+.135W_{t-2}$	$+.134W_{t-3}$	$+.068W_{t-12}$			.189	.00205
	(.00017)	(.042)	(.042)	(.041)	(.040)				
5.	.00057	$+.080\Delta_{t-1}$	$+.208\Delta_{t-2}$	$+.261\Delta_{t-12}$	$+.183W_{t-1}$	$+.069W_{t-2}$	$+.045W_{t-3}$	.327	.00187
	(.00019)	(.067)	(.066)	(.060)	(.040)	(.041)	(.040)		
6.	$-.00068$	$+.978R_t$						.291	.00196
	(.00030)	(.102)							
7.	$-.00046$	$+.775R_t$	$+ 133W_{t-1}$	$+.080W_{t-2}$	$+.073W_{t-3}$			.353	.00183
	(.00028)	(.103)	(.036)	(.036)	(.036)				
8.	$-.00044$	$+.571R_t$	$+.151W_{t-1}$	$+.065W_{t-2}$	$+.045W_{t-3}$	$+.095\Delta_{t-2}$	$+.204\Delta_{t-12}$	.390	.00178
	(00029)	(.120)	(037)	(.038)	(038)	(.067)	(.059)		
9.	$- 00047$	$+.593R_t$	$+.158W_{t-1}$			$+.136\Delta_{t-2}$	$+.207\Delta_{t-12}$	.383	.00179
	(.00029)	(.121)	(.038)			(.064)	(.059)		

Note Standard errors are shown in parentheses below estimates of regression coefficients. R^2 is the coefficient of determination, adjusted for degrees of freedom, $s(\epsilon)$ is the standard error of the regression residuals

TABLE 2B—AUTOCORRELATION STATISTICS FOR REGRESSION RESIDUALS

Regressions	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}	$s(\rho)$	BP
1.	−.02	− 05	−.08	.08	.07	.04	−.04	.14	.13	.05	.01	.20	.067	23
2.	−.01	−.01	−.13	.05	.08	−.08	−.03	.11	.12	.07	.00	.00	.069	14
3.	.18	.24	.08	.16	.16	.22	.10	.22	.19	.21	.16	.33	.067	104
4.	17	.22	.05	.12	.14	.17	.08	.22	.22	.18	.14	.29	.069	81
5.	−.04	−.02	−.13	−.01	.08	.09	−.07	.08	.12	.08	.01	.02	.069	14
6.	11	.12	−.02	−.01	−.02	−.02	−.07	.05	.10	.10	.03	.19	.067	21
7.	.01	.09	−.07	−.03	−.02	.02	−.11	.05	.05	.07	.03	.25	.067	22
8.	−.04	.01	−.16	−.06	.02	.04	−.11	.04	.08	.05	−.00	.06	.069	12
9.	−.02	−.00	−.14	−.06	.04	.05	−.11	.05	.08	.03	−.01	.06	.069	12

Note: The ρ_τ are Box-Jenkins estimates of the autocorrelations of the residuals; $s(\rho)$ is the approximate standard error of each autocorrelation estimate under the hypothesis that the true autocorrelation is zero; and BP is the Box-Pierce statistic for the first twelve residual autocorrelations.

do not dampen at higher order lags. Thus the nonstationarity of the *CPI* inflation rate, which is evident from the nondampening estimated autocorrelations of Δ_t in Table 1, remains in the residuals from the regressions of Δ_t on lagged *WPI* inflation rates.

The fifth regression in Table 2 models Δ_t in terms of Δ_{t-1}, Δ_{t-2}, the seasonal term Δ_{t-12}, and the first three lags of W_t. The second- and third-order lags of the *WPI* inflation rate, which are important in the third regression, are unimportant when included along with the lags of the *CPI* inflation rate in the fifth regression. On the other hand, the first-order lag of the *WPI* inflation rate remains strong in the fifth regression, while the first-order lag of the *CPI* inflation rate, which shows up clearly in the second regression, becomes unimportant. All coefficients of lagged inflation rates are attenuated somewhat when the interest rate R_t is in-

cluded in the regressions, but a picture similar to that described above nevertheless emerges.

From the general "jumpiness" of the coefficients in the Table 2 regressions, it seems clear that the lags of the *CPI* and *WPI* inflation rates are picking up similar lagged inflation effects. Moreover, taking the evidence at face value, the ninth regression contradicts my model. There are three lagged inflation rates W_{t-1}, Δ_{t-2}, and Δ_{t-12} whose implications for the expected value of $\tilde{\Delta}_t$ are apparently not fully incorporated into the interest rate R_t set at the end of month t − 1.

IV. Spurious Lags in Measured Inflation Rates

Since the simple regression of Δ_t on R_t is buried in the middle of Table 2, one might tend to overlook the fact that it competes fairly well with the models of the expected inflation rate that also include lagged inflation effects. The interest rate R_t set at the end of month t − 1 accounts for about 30 percent of the variance of the inflation rate for month t. After searching rather thoroughly through the data, one uncovers lagged inflation variables that absorb only an additional 10 percent of the variance of Δ_t and provide close to pure white regression residuals in place of those, from the regression of Δ_t on R_t, which have some slight hints of gray. Moreover, I now argue that the additional contribution of lagged inflation variables to the prediction of the inflation rate might reflect systematic measurement errors in estimates of inflation rates that are properly ignored by the market in setting interest rates.

Consider the way the *CPI* is constructed.[2] The food and fuel components, about 25-28 percent of the index, are priced monthly in all locations. Almost all components are also priced monthly in the five largest cities, and these cities account for a little more than 30 percent of the *CPI*. Making a rough correction for overlap, we can infer that about 50 percent of the prices included in the *CPI* are sampled monthly. Most other prices are collected quarterly, but on a rotating basis across locations so that there is some revision of prices each month. Prior to 1964 prices for nonsampled items were imputed, usually from sampled prices in the five largest cities. Since the 1964 revision of the index, prices for nonsampled items and locations are generally held constant at their most recently sampled values.

Thus, since 1964 about 15 percent of the monthly prices reported as changed, those for locations sampled during the month that are on a quarterly sampling interval, reflect changes for the preceding two months as well. Since the money prices of goods tend to move together, such lags in the collection process introduce spurious short-term autocorrelation in measured monthly inflation rates. Reinforcing these spurious short-term autocorrelations, prices of owner-occupied housing, about 6.3 percent of the total index, are included in the *CPI* as a three-month moving average.

There are also annual seasonals in the index that are to some extent spurious. Some items, for example, college textbooks, are only sampled annually. In constructing the rental component, about 5.5 percent of the total index, the rent on a given apartment is sampled every six months, with different apartments staggered across months. Most rental contracts are only renegotiated annually, however, so that changes in reported rents usually reflect actual changes in the rental market over a one-year period.

The information in past inflation rates which the market seems to neglect in setting interest rates is suspiciously similar to the spurious short-term lags and seasonals discussed above. The high performing ninth regression in Table 2 includes a significant seasonal as well as first-and second-order lagged inflation terms. The Nelson-Schwert optimal time-series predictor of inflation involves a first-order moving average term, and they recognize that predictions would be improved by including a seasonal. Hess

[2] A detailed description of the *CPI* is in the *BLS* reference.

and Bicksler include the seasonal and suggest that a second-order moving average term would further improve the predictive power of the optimal forecaster, all of which is in line with the quirks in the procedures used to calculate the *CPI*.

V. The Wanderings of the Expected Real Return

Although both null hypotheses, market efficiency and a constant expected real return, are surely to some extent false, of the two the constant expected real return hypothesis is on weaker theoretical ground. Theory implies that a well-functioning market is efficient in the sense that it correctly uses available information in forecasting the inflation rate. However, testing market efficiency, which is the stated goal of my paper, always requires a model of market equilibrium which in effect provides specific propositions about what the market is trying to do when it sets the prices of securities. The arbitrary hypothesis that the market sets Treasury Bill prices so that it perceives constant expected real returns plays this role in my work. It is one of many possible devices which would allow tests of efficiency to go forward; as a model of market equilibrium, it has little basis in economic theory. In fact, in my 1976 paper there is evidence that the expected real return on 1-month Treasury Bills is related to uncertainty about future expected inflation rates, which itself seems to wander noticeably during the sample period 1953–71. The estimated expected real return seems to follow a nonstationary process close to a random walk.

However, the fact that the expected real return is not literally constant is not sufficient to reject the hypothesis of a constant expected real return as a simplifying approximation for tests of market efficiency. The interesting implication of the joint hypotheses that the market is efficient and the expected real return is constant is that all variation in the nominal interest rate reflects variation in correctly assessed expected inflation rates. Given an efficient market, this remains an interesting approximate description of the world as long as variation in the expected real return is a small part of the variation in the nominal interest rate.

In my 1976 paper, I estimate that, like the expected real return, the expected inflation rate follows a process close to a random walk. But its random steps from one month to the next have a variance that swamps the variance of the month to month changes in the expected real return. The result is that, at least during the 1953–71 period, variation in the nominal interest rate seems to reflect almost entirely variation in the expected inflation rate. Nelson and Schwert obtain a variety of estimates of the variation in $E(\tilde{r})$, most of which are larger than mine, but their final sentence seems to agree with my basic point: "Our analysis does suggest that expectations of inflation have accounted for most of the variation in short-term interest rates during the postwar period, and that those expectations embody significant information beyond that contained in past inflation rates alone."

VI. Broadening the Attack on Efficiency

The tests in my 1975 paper focus on the information in past inflation rates. In assessing expected future inflation rates, an efficient or rational market correctly uses all available information. Let ϕ_{t-1} be the set of information available at $t-1$ and relevant for forecasting $\tilde{\Delta}_t$. Then the joint hypotheses that the market is efficient and that it sets bill prices so that expected real returns are constant can be expressed as

$$E(\tilde{\Delta}_t \mid R_t, \phi_{t-1}) = -E(\tilde{r}) + R_t \tag{6}$$

which is (4) but with the broader information set ϕ_{t-1} substituted for the time-series of past inflation rates.

In intuitive terms, although ϕ_{t-1} contains the information relevant for assessing the expected value of $\tilde{\Delta}_t$, equation (6) says that this information is summarized in the nominal interest rate R_t. Once R_t is set at $t-1$, the information in ϕ_{t-1} becomes redundant for assessing the expected value of $\tilde{\Delta}_t$. In statistical terms, equation (6) is a statement about the form of a regression function—the conditional expected value of $\tilde{\Delta}_t$ as a

TABLE 3A—ESTIMATED REGRESSIONS OF THE *CPI* INFLATION RATE Δ_t AND THE REAL RETURN r_t ON THE EMPLOYMENT TO POPULATION RATIO $(E/P)_{t-6}$ AND THE NOMINAL INTEREST RATE R_t SET AT THE END OF THE MONTH t − 1

Regressions				R^2	$s(\epsilon)$
1.	$\Delta_t = -.00068$	$+.978R_t$		.291	.00196
	(.00030)	(.102)			
2.	$\Delta_t = -.04851$	$+.641R_t$	$+.088(E/P)_{t-6}$	.357	.00187
	(.00988)	(119)	(.018)		
3.	$\Delta_t = -.07831$		$+.145(E/P)_{t-6}$	.276	.00198
	(.00868)		(.016)		
4.	$r_t = .00068$	$+.022R_t$		.000	.00196
	(.00030)	(.102)			
5.	$r_t = .03182$		$-.056(E/P)_{t-6}$	.055	.00190
	(.00832)		(.015)		
6.	$r_t = .04851$	$+.359R_t$	$-.088(E/P)_{t-6}$	.088	.00187
	(.00988)	(.119)	(.018)		

Note: See note to Table 2A.

TABLE 3B—AUTOCORRELATION STATISTICS FOR REGRESSION RESIDUALS

Regressions	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_7	ρ_8	ρ_9	ρ_{10}	ρ_{11}	ρ_{12}	$s(\rho)$	*BP*
1.	.11	.12	- 02	-.01	- 02	-.02	- 07	.05	.10	.10	.03	.19	.07	21
2	.04	.03	- 12	-.10	-.09	-.08	- 14	.02	.06	.06	-.02	.17	.07	22
3.	.16	.12	-.01	.03	.03	.04	-.01	.13	.15	.14	.05	.21	.07	33
4	.11	12	-.02	-.01	-.02	-.02	-.07	.05	.10	.10	.03	19	.07	21
5.	.07	.07	- 08	-.07	- 07	-.07	- 12	02	07	.08	.00	.18	.07	20
6	.04	03	- 12	-.10	-.09	-.08	-.14	02	.06	.06	- 02	.17	.07	22

Note See note to Table 2B.

function of R_t and variables in ϕ_{t-1}. It implies that in any estimated regression of $\tilde{\Delta}_t$ on R_t and any other variables whose values are known at t − 1, the coefficient of R_t should be statistically indistinguishable from 1.0, the coefficients on other included variables should be indistinguishable from zero, and the regression residuals should be serially uncorrelated.

John Carlson examines market efficiency in this broader context. His candidate for a noninflation variable that can be used to improve predictions of inflation is $(E/P)_{t-6}$, the ratio of seasonally adjusted employment to population, measured six months before the month in which Δ_t is observed. My replication of his regression of Δ_t on R_t and $(E/P)_{t-6}$ for the 1953–71 period is the second regression in Table 3.[3] Comparing this regression with the regression of Δ_t on R_t, the first regression in the table, one finds that adding $(E/P)_{t-6}$ to the model reduces the residual standard error by 5 percent. The contribution of $(E/P)_{t-6}$ is nevertheless beyond standard sampling limits, and including it in the equation pushes the coefficient of R_t well below the value of unity called for by my model. Thus, there is some interest in tracing how $(E/P)_{t-6}$ does its work.

I have subjected the employment to population ratio to the same sort of attack (that is, including lags of the *CPI* in the multiple regression) leveled against the lags of the *WPI* in Table 2. Unlike the lags of the *WPI*, $(E/P)_{t-6}$ passes the data-dredging gantlet with its coefficient and statistical significance unscathed, indicating that this variable is not just proxying for short-term lagged inflation effects. Moreover, the third

[3] I thank Carlson for making his data available.

regression in Table 3 indicates that a model for the expected inflation rate based on $(E/P)_{t-6}$ alone does almost as well, in terms of coefficient of determination, as the interest rate model, although it is important to note that the interest rate alone produces residual autocorrelations that are closer to zero. Finally, there is nothing magical about the sixth lag of the employment to population ratio. In replications of the second and third regressions in Table 3 with other short-term lags and even leads of the variable, the results are similar to those obtained with $(E/P)_{t-6}$.

Some of these findings are a consequence of the fact that R_t and $(E/P)_{t-6}$ are mean nonstationary series whose levels move pretty much together through time. The correlation between the levels of the two series is about .6. From Table 1 we can see that the levels of the two series have similar autocorrelations, with individual autocorrelations close to one; and the autocorrelations of the monthly changes in the two series are similar. However, only the general movements of the two series are similar. A regression of the level of either on that of the other for any given short-term lead or lag between the series produces high residual autocorrelations—on the order of those shown for the levels of the series in Table 1. The correlations between different leads and lags of the monthly changes of the two series are low—on the order of .15 or less.

Given that the interest rate tracks the expected inflation rate, and given that the interest rate and the employment to population ratio are correlated, it is not surprising that $(E/P)_{t-6}$ likewise appears to track the expected inflation rate. But none of this explains why, contrary to my model, the employment to population ratio seems to make a contribution to forecasts of inflation beyond that provided by the interest rate alone. I suggest two possibilities, both of which might contain some amount of truth.

First, there are systematic measurement errors in the *CPI* in addition to those caused by the lags in price collection discussed earlier. There is some amount of quality change in the items in the *CPI* which shows up spuriously as price change. Since it represents an average across items, the spurious component of the measured inflation rate which actually represents quality change might end up as a nonstationary variable whose slow wandering through time is spuriously correlated with other nonstationary series like the employment to population ratio.[4] Thus some part of the marginal contribution of the employment to population ratio to the description of the measured inflation rate might represent a spurious component of the inflation rate which is properly ignored by the market in setting the interest rate.

Second, we have spent much space discussing measurement errors in the *CPI*. The interest rates also contain measurement errors. The rates are estimated as averages of end-of-day bid and asked prices quoted by a particular dealer, Salomon Brothers. The prices quoted at the end of the day by dealers in this market are not actual trading prices or prices at which they would trade, but rather they are what the dealer feels to be the state of the market; and there is some dispersion in the quotes on the same bill given by different dealers. When there is measurement error in an explanatory variable, that variable's coefficient is likely to be spuriously attenuated when another variable with which it is correlated is also included as an explanatory variable in a regression.[5] This might explain to some extent why the estimated coefficient of the interest rate in Table 3 declines when the employment to population variable is added to the inflation regression.

Since the interest rate is also correlated with lagged inflation rates and with time-series forecasts of the inflation rate based on lagged inflation rates, the same phenomenon might explain the decline in the coefficient of the interest rate that occurs when

[4]The levels of series that are mean nonstationary are likely to be correlated, although not necessarily positively, even when the series are generated independently. This problem was originally discussed by G. Udny Yule. A more recent treatment is given by C. W. J. Granger and P. Newbold.

[5]The problem is discussed and illustrated by Merton H. Miller and Myron Scholes.

different versions of lagged inflation variables are included in the regression of the inflation rate on the interest rate. In this respect, it is interesting that the decline in the coefficient of the interest rate that occurs when $(E/P)_{t-6}$ is added to the regression is similar in magnitude to the declines observed in Table 2 when different versions of lagged inflation variables are included.

VII. The Expected Real Return and $(E/P)_{t-6}$

There is another way to look at the issues raised by the employment to population ratio that may be helpful. The joint hypotheses that the market is efficient and that it sets bill prices so that expected real returns are constant through time imply

$$E(\tilde{r}_t \mid \phi_{t-1}) = E(\tilde{r}) \quad (7)$$

In words, the model says that the market prices Treasury Bills so that the expected real return on a 1-month bill from the end of any month t − 1 to the end of month t is the constant $E(\tilde{r})$ which is unrelated to values of information variables available at t − 1. Equation (7) is a statement about the form of a regression function, the expected value of $\tilde{r}_t$ as a function of information variables known at t − 1. It says that in any estimated regression of the real return r_t on values of variables known at t − 1, the estimated regression coefficients of these variables should be statistically indistinguishable from zero.

The last three equations in Table 3 estimate the regression of r_t first on the interest rate R_t set at t − 1, then on the employment to population ratio $(E/P)_{t-6}$, and finally on the two variables together. The results are fascinating. The fourth regression says that there is no relationship between r_t and R_t alone. Thus, any variation in the underlying expected real return shows no relationship to variation in the interest rate alone. There is, however, a negative relationship between r_t and $(E/P)_{t-6}$. The expected real return apparently varies inversely with the employment to population ratio, a result which at least some macroeconomists would find curious. Moreover, when R_t is included in the regression along with $(E/P)_{t-6}$, the coefficient of the latter becomes more negative, and the interest rate now has a significantly positive coefficient.

There are at least three explanations of these phenomena, and they are not mutually exclusive. First, we may be measuring variation in the equilibrium expected real return. It is curious, however, that the variation in the expected real return does not show up directly in the nominal interest rate. (That, after all, is the implication of the zero correlation between r_t and R_t.) Second, we may be observing a nonstationarity in the true inflation rate which is overlooked by the market in setting the interest rate; that is, a market inefficiency has been uncovered. Finally, we may be witnessing a nonstationarity in the measured inflation rate which is a manifestation of measurement error (the spurious effects of quality changes suggest themselves), in which case the market is to be congratulated for ignoring this component of the measured inflation rate in setting the interest rate.

There is no way to choose unambiguously among these three interpretations of the real return regressions, and all of them might contain some truth. The important point, I think, is that the component of the real return or of the inflation rate which is correctly or incorrectly overlooked by the market in setting the interest rate is a small part of the variability of either the real return or the inflation rate. Thus, from equations (1) and (4), the forecasting error of my model for any month t is $-[\tilde{r}_t - E(\tilde{r})]$, the negative of the deviation of the real return from its assumed constant expected value. The coefficient of determination in the estimated regression of the real return r_t on $(E/P)_{t-6}$ in Table 3 is .055; less than 6 percent of the variance of the forecasting errors of my model can be explained by $(E/P)_{t-6}$. Although the employment to population ratio raises interesting and largely unresolved questions, we are nevertheless left with the impression that the interest rate captures the largest part of what is predictable in the inflation rate.

VIII. Conclusions

The interest in my work taken by scholars such as Bicksler, Carlson, Hess, Joines, and Nelson and Schwert is gratifying. However, the challenges of these authors do not imply the joint hypotheses that the Treasury Bill market is efficient and that the market sets prices so that expected real returns are constant through time should be rejected as a useful description of the Treasury Bill market for the 1953-71 period.

Taking the data at face value, the interest rate remains the best single predictor of the inflation rate; and nobody has uncovered variables that make substantial contributions to the prediction of inflation beyond that provided by the interest rate alone. Moreover, one of the more interesting propositions of the model, that the largest part of the variation in nominal interest rates reflects variation in expected inflation rates, seems intact. Finally, although the model is not an exact description of the world, the specific deviations discovered so far are to some extent manifestations of measurement errors in the estimates of inflation rates and interest rates.

REFERENCES

George P. Box and Gwilym M. Jenkins, *Time Series Analysis*, San Francisco 1970.

J. A. Carlson, "Short-Term Interest Rates as Predictors of Inflation: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 469-75.

E. F. Fama, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.* June 1975, *65*, 269-82.

——, "Inflation Uncertainty and Expected Returns on Treasury Bills," *J Polit. Econ.*, June 1976, *84*, 427-48.

C. W. J. Granger and P. Newbold, "Spurious Regressions in Econometrics," *J. Econometrics*, July 1974, *2*, 111-20.

P. J. Hess and J. L. Bicksler, "Capital Asset Prices versus Time Series Models as Predictors of Inflation: The Expected Real Rate of Interest and Market Efficiency," *J. Finance Econ.*, Dec. 1975, *2*, 341-60.

D. Joines, "Short-Term Interest Rates as Predictors of Inflation: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 476-77.

M. H. Miller and M. Scholes, "Rates of Return in Relation to Risk: A Reexamination of Some Recent Findings," in Michael Jensen, ed., *Studies in the Theory of Capital Markets*, New York 1972.

C. R. Nelson and G. W. Schwert, "Short-Term Interest Rates as Predictors of Inflation: On Testing the Hypothesis That the Real Rate of Interest Is Constant," *Amer. Econ. Rev.*, June 1977, *67*, 478-86.

G. Udny Yule, "Why Do We Sometimes Get Nonsense Correlations between Time Series? A Study in Sampling and the Nature of Time Series," *J Royal Statist Soc.*, Part 1, Jan. 1926, *89*, 1-69.

U.S. Bureau of Labor Statistics (*BLS*), *The Consumer Price Index: History and Techniques*, Bull. 1517, Washington 1966.

The Measurement and Trend of Inequality: Comment

By ERIC R. NELSON*

In a recent article in this *Review*, Morton Paglin proposes a measure of concentration which he claims will correct the Gini ratio for overstatement in populations where each individual may be expected to vary his income or wealth over his life cycle. The Paglin measure suggests that the Gini ratio overstates income concentration in the United States by a third and does not reveal a 23 percent decline in concentration which has occurred between 1947 and 1972. These results are based on an improper decomposition of the Gini ratio.

In this note I decompose the Gini ratio to derive the Paglin measure. It is shown to correct for neither demographic nor life cycle income effects and to understate intracohort income or wealth inequality.[1] If properly normalized, the Paglin measure shows that the Gini overstates *U.S.* income concentration by 7 percent or less, and indicates a decline in *U.S.* interfamily income concentration of less than 12 percent. Four percentage points of this decline are the result of demographic change, while less than 7 points (comparable to the 5 percent decline of the Gini) are the result of decreased income concentration within cohorts of family heads. Decomposition further reveals major changes in the division of income among cohorts and in concentration of income within cohorts. These changes compensate each other in the Paglin measure, the use of which as a measure of income or wealth concentration is not recommended.

The Gini concentration ratio

$$(1) \qquad G = \frac{1}{2n\Sigma x_i}\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|$$

can be disaggregated in the case of grouped data into measures of intragroup concentration and between-group concentration[2]

(2)

$$G = \sum_{i=1}^{Q} \lambda_i \gamma_i G_i + \frac{1}{2n^2\mu}\sum_{\substack{i\in I}}\sum_{\substack{j\in J \\ I\neq J}} |x_i - x_j|$$

where G_i is the intragroup Gini, μ represents mean (of society or of a group), λ_i is the group's numerical importance in the population (n_i/n), and γ_i is its share of income (or wealth) $n_i\mu_i/n\mu$. In the limiting case where families are grouped by income magnitude, so group distributions cannot overlap, then

(3)

$$G = \sum_{i=1}^{Q} \lambda_i \gamma_i G_i + \frac{1}{2n^2\mu}\sum_{i=1}^{Q}\sum_{j=1}^{Q} n_i n_j |\mu_i - \mu_j|$$

where the second term is identical to Paglin's age-Gini.

Income distributions by cohort overlap extensively, so Paglin's use of (3) underestimates the absolute age-income correction term[3] and thus overestimates his *P*-

*Centre Ivoirien de Recherches Economiques et Sociales, Abidjan, and Center for Research on Economic Development, University of Michigan. I am indebted to a referee who discovered my use of an incorrect data set in an earlier draft

[1]As measured by the Gini ratio, which may not be a reliable welfare measure. See A. B Atkinson, John Fei and Gary Fields, and Michael Rothschild and Joseph Stiglitz, among others. Graham Pyatt demonstrates that the Gini can be interpreted as "the average gain to be expected if each individual has the choice of being [himself] or some other member of the population drawn at random, expressed as a proportion of the average level of income" (p. 4). The Gini is then a legitimate welfare measure only within a group of comparable individuals. A cohort may be such a group; Paglin correctly suggests that the whole population is not The welfare interpretation of the Paglin measure is obscure.

[2]Similar manipulation is suggested by N. Bhattacharya and B. Mahalonobis.

[3]$\Sigma|x| \geq |\Sigma x|$, with equality if and only if $x \geq 0$ for all x.

measure, which becomes

$$(4) \qquad P = \sum_{i=1}^{Q} \lambda_i \gamma_i G_i$$

This is the sum of the products of three terms: the intracohort Gini ratio, the cohort's size as a fraction of the total population, and its total income share. Even in the most favorable case where cohort income distributions do not overlap, P is standardized on neither the demographic structure nor the life cycle income profile of the population and cannot be recommended as a measure of inequality[4] or its change.[5]

The Paglin measure is bounded above by the Gini rather than by unity; it is not homogeneous of degree zero in the demographic and age-income profile of the population. If the P-measure is normalized by division by $\Sigma\, \lambda_i \gamma_i$ it indicates that the Gini overstates *U.S.* family income concentration by 7.5 percent for 1947 and by 8.6 percent for 1972; standardizing on age- or income-profile weights or equal weights, as I have done in Table 1, reduces this overstatement.[6]

The change in the Paglin measure over time can be decomposed by total differentiation of (4), yielding

$$(5) \quad dP = \Sigma\, \lambda_i \gamma_i G_i \left(\frac{d\lambda_i}{\lambda_i} + \frac{d\gamma_i}{\gamma_i} + \frac{dG_i}{G_i}\right)$$

Estimation of the magnitude of these terms does not support the hypothesis that a considerable decline in inequality occurred between 1947 and 1972. In Table 2 the Gini ratio is disaggregated by factor, and the true factors are compared to the Paglin model estimates. In Table 3 the change in the Paglin measure from 1947 to 1972 is calculated on the basis of this information. Equal weights, *log* differences, or standardization on either year's population or age-income profile, as summarized in Table 1, do not support the Paglin thesis. Laspeyre standardization reveals a decrease of only 6.6 percent in the Gini between 1947 and 1972. This decline is comparable to that of the Gini.

TABLE 1—COMPARISON OF THE GINI RATIO AND WEIGHTED AVERAGES OF INTRACOHORT GINIS, *U.S.* FAMILIES, 1947 AND 1972

	1947	1972	Change (percent)
Paglin Measure (P)	0.303	0.239	-23.7
Published Gini (G)	0.378	0.359	-5.2
Intracohort Ginis			
Unweighted average	0.360	0.342	-5.2
1947 Weights:			
Population (λ)	0.358	0.334	-6.6
Income share (γ)	0.356	0.332	-6.6
1972 Weights			
Population (λ)	0.359	0.337	-6.3
Income share (γ)	0.355	0.333	-6.4

Source See Table 2 and Paglin.

To study the change in the Gini ratio from 1947 to 1972 I have used Census Bureau statistics on the distribution of families by income class and by age of family head for the two endpoint years. Within cohorts it is legitimate to use equation (3) rather than (2), since incomes are monotonically ordered by income grouping. The intracell G_i are assumed to be zero for each cohort-income class,[7] and the geometric mean of cell endpoints approximates cell mean incomes.[8] Cohort mean incomes for 1947

[4]The changing demographic composition of families further complicates the use of such a measure. Alice Rivlin provides an overview of the problems inherent in the study of measures of inequality among cohorts of heads of families or households and in the interpretation of changes in such measures over time.

[5]The Gini ratio may be decomposed for the distribution of wealth, for which Paglin claims a P-measure of 0.50 for 1962 (versus a Gini of 0.76). In a study which followed population cohorts from 1947 to 1962—and which found an insignificant decrease in wealth inequality within cohorts—John B. Lansing and John Sunquist calculate intracohort wealth Ginis between 0.62 and 0.67 in 1962. This provides further supporting evidence that Paglin has understated the degree of wealth concentration in the United States.

[6]When calculated from equation (2), the "true" value of P is 0.069 in 1947 and 0.061 in 1972. If the measure had been correctly derived the need for normalization to the range 0–1 would have been clear.

[7]This leads to a 5-point (1.3 percent) computation error for the overall Gini for 1947, and a 6-point (1.8 percent) error for 1972. Paglin age-Ginis were underestimated by 1 point (1.2 percent) and 3 points (2.8 percent), respectively.

[8]Use of the geometric mean led to an inconsistent estimate for the mean of the open interval (over $10,000 in 1947, $50,000 in 1972) in most cohorts. Approximate estimates were then used to close the interval and the distributions were reestimated.

TABLE 2—DISAGGREGATION OF THE GINI RATIO BY FACTOR

	Age of Family Head:					
	14-24	25-34	35-44	45-54	55-64	65+
1. Cohort share λ						
a. 1947	.0490	.2173	.2378	.2137	.1641	.1170
b. 1972	.0771	.2196	.1972	.2071	.1594	.1359
2. Income share γ						
a. 1947	.0359	.1978	.2533	.2419	.1826	.0885
b. 1972	.0482	.2035	.2249	.2573	.1737	.0924
3. Cohort Gini G_i						
a. 1947	.2759	.3033	.3363	.3522	.3873	.5055
b 1972	.3340	.3027	.3140	.3224	.3612	.4169
4. Total $\lambda\ \gamma\ G_i$						
a. 1947	.0005	.0131	.0203	.0182	.0116	.0052
b. 1972	.0012	.0135	.0139	.0172	.0100	.0054
5. Means μ_i/μ						
a. 1947	.7328	.9060	1.0654	1.1323	1.1126	.7547
b. 1972	.6251	.9267	1.1402	1.2428	1.0897	.6619
	Interaction Term With Cohort of: 1947			1972		
6. 14-24						
a. True		.0144	.0186	.0225	0143	.0057
b. Paglin		.0051	.0079	.0099	.0057	.0004
7. 25 34·						
a True	0072		.0379	.0473	.0298	.0111
b Paglin	.0019		.0092	.0144	.0057	.0081
8 35 44.						
a. True	.0111	.0419		.0353	.0222	.0081
b. Paglin	.0039	.0083		.0042	.0016	.0132
9 45 54:						
a. True	.0113	.0433	.0420		.0216	.0079
b Paglin	.0042	.0106	.0034		.0051	.0168
10 55-64:						
a. True	.0087	.0335	.0326	.0282		.0075
b. Paglin	.0031	.0074	.0018	0007		.0095
11 65+						
a True	.0037	.0142	.0140	.0121	.0100	
b. Paglin	.0001	.0039	.0087	.0095	.0069	

Source U.S. Bureau of the Census: No. 5, Table 5; and No. 87, Table 3.

TABLE 3—CHANGE IN THE PAGLIN MEASURE BY COMPONENT, 1947 TO 1972

	Age of Family Head:						Average Weighted by:[a]		
	14-24	25-34	35-44	45-54	55-64	65+	Population	Income	Total
Percentage Change									
1 Cohort	57.3	0.6	-17.1	-3.1	-2.9	19.1	-3.9	-4.8	-4.3
2 Income	34.2	2.9	-11.2	6.4	-4.9	4.5	-1.3	-1.5	-1.3
3 Gini	21.1	-0.2	-6.6	-8.5	-6.7	-17.5	-6.6	-6.6	-6.5
4. Total	155.5	3.3	-31.3	-5.7	-13.8	2.6	-13.1	-13.7	-14.8
5 Means	-14.7	2.2	7.0	9.8	-2.1	-12.3			
Interactions:									
6 True	79.6	0.2	-13.8	-1.7	-1.7	25.5			
7 Paglin	120.7	33.2	38.3	77.7	75.4	65.5			

Source See Table 2.

[a]Weights are Laspeyre weighting based on 1947 population. Use of 1972 weights or difference of logarithms does not lead to a substantial difference.

were estimated using the median/mean ratio for 1972, when both statistics are available.[9] The elements of Table 2 correspond to those of equation (2), with the exception of the Paglin interaction (rows 6–11(b)) which are calculated from (3), and the cohort mean incomes (row 5, equal to row 2 divided by row 1).

It is obvious from rows 6–11 of Table 2 that the Paglin age-Gini interaction terms underestimate the true income overlap between cohorts by as much as 98 percent (row 10, col. 4, or row 11, col. 2) for 1947. Paglin's version of P is thus a biased version of the true P in (4), as is clear from the elements of P in row 4. These terms are more than one order of magnitude less than the Gini, and would not have been recommended as a measure of "perfect inequality" comparable to the Gini. Although intracohort Ginis (row 3) are independent of the age-income and demographic profiles of the population, they are of the same size as the overall Gini and give only modest support to the Paglin claim that the Gini overstates inequality: in Table 1, various weighted sums of these Ginis "correct" the published Gini by 4 to 7 percent, rather than the one-third claimed by Paglin.

The change in the components of the Gini, shown in Table 2 and summarized in Table 3, show that the Paglin measure—like any index—can conceal large compensating changes in its components. Minor increases in inequality among the cohorts most well off can compensate for major decreases in inequality among the poorer cohorts. Unlike other indices, however, "goods" and "bads" can compensate in the Paglin measure. For instance, the mean income of the oldest cohort has not kept up with the population mean income (Table 3, row 5) while intracohort inequality has decreased (row 3); the aging of the population has made these factors compensate in the Paglin measure, which evaluates as a good the larger percentage of more equally poor old persons. It is also clear that the P-measure does not standardize on the age and income structure, since although the change in intracohort inequality is the largest single determinant of the change in P (row 3), demographic effects are nearly as important (row 1) and changes in the age-income profile are not inconsequential (row 2). While the inequality factor in equation (5) declined 6.5 percent, the cohort weights caused a 3.9–4.8 percent decline. Last, the changes in the interaction terms, reflecting an equalization effect, move in the opposite direction from changes in the true interaction for the cohorts 35–64 (row 7, as compared to row 6 for the true). A large part of the "true" Paglin measure consists of overlap between cohort income distributions, and the change in this interaction over time is much greater than changes in inequality within cohorts.

Morton Paglin has claimed that by his adjustment of the Gini ratio for the age-income profile of the population:

> ... we not only avoid cumbersome data problems, but we open up to explicit examination the meaning of perfect equality. We also separate the issue of *intra*family variation in income over the life cycle (accounting for one-third of Lorenzian inequality) from the more basic issue of *inter*family differences in lifetime incomes. An application of the new concepts to *U.S.* income and wealth data reveals that estimates of inequality have been overstated by 50 percent, and the trend of inequality from 1947 to 1972 has declined by 23 percent. [p. 608]

I have demonstrated that the Paglin measure does not isolate the issue of intrafamily income variation from that of interfamily differences even in the ideal case estimated by Paglin; in reality this measure understates inequality seriously where cohort income distributions overlap. Last, investigation of the factors influencing the Paglin measure provides no substantiation of his claim that inequality decreased from

[9] Under reasonable assumptions such as lognormality of the income distribution, this will overstate γ_i and thus overestimate the P-measure for 1947. The decline in the P-measure shown in Tables 2 and 3 thus overstates the true decline and provides a strong test for our contention that Paglin has overestimated this decline.

1947 to 1972. Paglin has correctly noted that the traditional Lorenz measure of concentration overstates inequality, but he has not provided an acceptable alternative index.

REFERENCES

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

N. Bhattacharya and B. Mahalanobis, "Regional Disparities in Household Consumption in India," *J. Amer. Statist. Assn.*, Mar. 1967, *62*, 141–63.

J. C. H. Fei and G. S. Fields, "The Indexability of Ordinal Measures of Inequality," disc. pap. no. 205, Yale Econ. Growth Center, New Haven 1974.

J. B. Lansing and J. Sunquist, "Cohort Analysis of Changes in the Distribution of Wealth," in Lee Soltow, ed., *Six Papers on the Size Distribution of Wealth and Income*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 33, New York 1969.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598 609.

G. Pyatt, "On the Interpretation and Disaggregation of Gini Coefficients," unpub. mimeo, Develop. Res. Center, IBRD, Washington 1975.

A. M. Rivlin, "Income Distribution—Can Economists Help?" *Amer. Econ. Rev. Proc.*, May 1975, *65*, 1–15.

M. Rothschild and J. E. Stiglitz, "Some Further Results on the Measurement of Inequality," *J. Econ. Theory*, Apr. 1973, *6*, 188–204.

U.S. Bureau of the Census, *Current Population Reports*, Series P-60, no. 5, 87, Washington 1948, 1973.

The Measurement and Trend of Inequality: Comment

By WILLIAM R. JOHNSON*

In a recent article in this *Review*, Morton Paglin proposes a remedy for the well-known failure of measures of current income inequality to account for life cycle effects. Paglin's approach is to subtract from the conventional Gini coefficient of current income inequality an "age-Gini," which represents the inequality of average income across age levels. The result, dubbed the "Paglin-Gini," is intended to measure the extent of permanent lifetime income inequality in a stationary economy.[1] Paglin then shows that his measure of lifetime income inequality has fallen in the recent past in the United States despite the near constancy of the Gini coefficient as conventionally measured. Paglin also constructs a Lorenz curve of the interage distribution of income and argues that divergences of current income from this Lorenz curve should be the true measure of permanent income inequality. Using this age-adjusted Lorenz curve, he then concludes that the share of the bottom fifth of the permanent income distribution has increased.

The purpose of this comment is to show, first, that Paglin's adjusted Gini coefficient will always underestimate the true extent of lifetime income inequality. Secondly, given the actual pattern of age effects on earnings in the United States, Paglin's estimate of the progress in the relative income share of the bottom of the income distribution will be overstated.

Although Paglin advances no formal model of income determination, I propose a very simple model which, I believe, captures the essence of Paglin's thinking. Let there be six equal-sized population groups with three age levels (subscripted by k) and two permanent income levels (subscripted by i). Current income for each cell, Y_{ik}, is the sum of permanent income and the age effect:

$$(1) \qquad Y_{ik} = P_i + A_{ik} \qquad \begin{matrix} i = 1,2 \\ k = 1,2,3 \end{matrix}$$

where P_i is the permanent income of group i and A_{ik} is the age effect of group ik. The effect of age on income is zero over the lifetime:

$$(2) \qquad \sum_{k=1}^{3} A_{ik} = 0 \qquad i = 1,2$$

Let us first consider the case in which age effects are independent of permanent income:

$$(3) \qquad Y_{ik} = P_i + A_k \qquad \begin{matrix} i = 1,2 \\ k = 1,2,3 \end{matrix}$$

It can be shown that the Paglin-Gini is always less than or equal to the true inequality of permanent income. The proof is straightforward. For a discrete distribution, the Gini coefficient of current incomes (Paglin calls it the Lorenz-Gini) is

$$(4) \qquad G_L = K \sum_{i=1}^{2} \sum_{k=1}^{3} \sum_{m=1}^{2} \sum_{j=1}^{3} | Y_{ik} - Y_{mj} |$$

where K is a term which depends on the size and the mean income of the population. Calculating (4) for incomes given by (3), we derive

$$(5) \qquad G_L = 2K \Big[3 | P_1 - P_2 | + 2 | A_1 - A_2 | + 2 | A_2 - A_3 | + 2 | A_3 - A_1 | + \sum_{\substack{j=1 \\ j \neq k}}^{3} \sum_{k=1}^{3} | P_1 - P_2 + A_j - A_k | \Big]$$

Similar calculations can derive the age-Gini

*University of Virginia.

[1] As Paglin points out, this will not be the measure of income inequality we would obtain if lifetime incomes were known precisely for each person, because of economic growth.

(G_A) and the Gini of permanent income (G_P):

$$(6)\quad G_A = 2K[4|A_1 - A_2| + 4|A_2 - A_3| + 4|A_3 - A_1|]$$

$$(7)\quad G_P = 2K[9|P_1 - P_2|]$$

It is clear that Paglin's measure ($G_L - G_A$) is less than G_P if, and only if,

$$(8)\quad -2|A_1 - A_2| - 2|A_2 - A_3| - 2|A_3 - A_1| + \sum_{j=1}^{3}\sum_{\substack{k=1\\ j\neq k}}^{3} |P_1 - P_2 + A_j - A_k| < 6|P_1 - P_2|$$

By the well-known property of absolute values that $|a + b| \leq |a| + |b|$, we know that

$$(9)\quad \sum_{j=1}^{3}\sum_{\substack{k=1\\ j\neq k}}^{3} |P_1 - P_2 + A_j - A_k| \leq 6|P_1 - P_2| + 2|A_1 - A_2| + 2|A_2 - A_3| + 2|A_3 - A_1|$$

Therefore, as long as the age effects are not all zero, the strict inequality in (8) holds, and the Paglin-Gini understates permanent income inequality.

We now turn to the situation in which age effects are not independent of perma- ment income. Although the bias in the overall Paglin-Gini is not necessarily greater than in the model described by (3), Paglin's estimate of the relative share of the bottom of the income distribution will likely be biassed upward. For simplicity, consider the six-cell model when age affects only the high permanent income group:

$$(10)\quad \begin{aligned} Y_{1k} &= P_1 \qquad k = 1,2,3\\ Y_{2k} &= P_2 + A_k\\ P_1 &< P_2 \end{aligned}$$

This is a simplification of the data Paglin presents in Figure 2a (p. 600) in which age effects are clearly greatest for those with the greatest permanent income. The theory of human capital and on-the-job training would also predict such a result. Three other assumptions are also consistent with the data:

$$(11)\quad \begin{aligned} A_1 &< 0 < A_2\\ A_1 &< A_3\\ Y_{21} &> Y_{11} \end{aligned}$$

That is, the age effect for the youngest age group is most negative, yet current income for this group (young high permanent income) is still greater than that for the young low permanent income group.

Consider first the ideal measure of divergence between the actual income share of the bottom of the permanent income distribution and the share under permanent income equality. This measure, the one we would use if we could observe permanent incomes, for the bottom sixth of the permanent income distribution in our example is simply:

$$(12)\quad (\bar{P} - P_1)/6\bar{P}$$

where $\bar{P} = (P_1 + P_2)/2$

Paglin's measure, on the other hand, subtracts the actual share of the lowest sixth of the current income distribution from the share of the bottom in the interage group distribution:

$$(13)\quad (\bar{P} + A_1/2 - P_1)/6\bar{P}$$

Clearly Paglin's measure of divergence (13) is less than the true measure (12) if and only if $A_1 < 0$, which is true by assumption. Only if age effects are identical for all permanent income groups will Paglin's measure not be biassed.

The intuitive reason for this result is that by assuming that all income groups experience the same age effect, Paglin implicitly overestimates the lifetime income of young (and old) low-income workers. By ascribing the *average* age income pattern to workers with the lowest incomes, Paglin understates the permanence of their low income status. Furthermore, the biasses become greater the more young and old workers there are in the population.

Paglin blames recent increases in the proportion of young and old in the population for the failure of the share of the bottom of

the current income distribution to reflect the (hypothesized) increase in the share of the bottom of the permanent income distribution. However, the data do not necessarily bear out this conclusion. For example, between 1969 and 1974, the fraction of total households headed by a person 14 to 24 years old or 65 years and older, rose from 26.3 percent to 28.3 percent. At the same time, the proportion of these households among households in the bottom fifth of the money income distribution fell from 59.6 percent to 57.0 percent.[2] Thus, while it is clear that the young and old constitute a disproportionate number of low income workers, it does not necessarily follow that the apparent stability of income inequality is attributable to the changing age composition of the population.[3] Likewise, while measures of current income inequality clearly overstate the degree of permanent income inequality, Paglin's adjustments are likely to understate lifetime inequality of incomes.

[2]See U S Bureau of the Census, p 5

[3]It is, of course, partly attributable to another demographic trend, the rise of the female-headed household

REFERENCES

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer Econ. Rev.*, Sept. 1975, *65*, 598-609.

U.S. Bureau of the Census, *Current Population Reports*, Series P-60, *Household Money Income in 1974 and Selected Social and Economic Characteristics of Households*, No. 100, Washington 1975.

The Measurement and Trend of Inequality: Comment

By SHELDON DANZIGER, ROBERT HAVEMAN, AND EUGENE SMOLENSKY*

In the September 1975 issue of this *Review*, Morton Paglin posits a new summary statistic, the "Paglin-Gini," which he uses to measure and analyze the level and trend of income inequality in the United States. This measure intends to recognize only income differences among families unrelated to the observed age-income profile as contributing to meaningful income inequality. This is to be contrasted with the standard Lorenz-Gini, in which all differences in income among living units contribute to inequality. Paglin's measure indicates substantially less inequality in the United States than is indicated by the Lorenz-Gini. Moreover, Paglin-inequality falls considerably in the post-World War II period while other measures indicate very little change in inequality over this period.

Paglin is addressing an important problem. For the postwar period when the age distribution, the length of time spent in school, and the propensity of the young and the old to form their own households have all changed rapidly, the trend in the conventional Gini coefficient has no obvious normative interpretation. However, it is often given one. Unfortunately, Paglin's proposed index of inequality does not meet its objective and, even if it did, its normative content would be no clearer than that of the standard measure. Indeed, no single indicator is sufficient to capture trends in normatively relevant inequality without further analysis. Like any other time-series, a time-series of inequality must be approached with a well-specified, multivariate model.

After describing Paglin's procedure in the next section, we challenge his measure in Section II. In Section III, Paglin's interpretation of policy-relevant inequality is questioned, and in Section IV the implications of the Paglin-Gini for the equity of the transfer system are analyzed. Section V is a summary of our critique.

I

The Paglin-Gini can be calculated for any size distribution of income, if the relationship between age and income is known from the same data. By ranking age cohorts from lowest to highest mean income and cumulating the percentage of observations (for example, families, households) and of income, a Lorenz-type curve is obtained which Paglin calls the *P*-reference line. The deviation of this curve from the 45 degree line of equality is to depict the inequality in incomes attributable to the relationship between age and income. This inequality is measured by a Gini coefficient which Paglin calls the age-Gini. The age-Gini measures the inequality which would exist if there were no variance around the mean income of each cohort, but there were differences among the means. Deducting the age-Gini from the standard Lorenz-Gini yields the Paglin-Gini. The Paglin-Gini is intended to measure the inequality attributable to the variation around the mean income of each cohort.

Paglin's purpose is to partition the area between the 45 degree line and the Lorenz curve into two parts: that inequality which to him is economically functional and, hence, of no concern for public policy, and the remaining inequality which one may or

*Institute for Research on Poverty, University of Wisconsin Madison. We gratefully acknowledge the support of funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Department of Health, Education and Welfare pursuant to the provision of the Economic Opportunity Act of 1964. Support was also provided by the Netherlands Institute for Advanced Study. We want to thank David Betson, Katharine Bradbury, and Alan Cohen for valuable suggestions. The opinions expressed are solely our own.

TABLE 1—LORENZ- AND PAGLIN-GINI COEFFICIENTS, 1965 AND 1972

	Actual 1972 (1)	1972 Gini Holding Constant at the 1965 Level: Cohort-Specific Gini Coefficients (2)	Distribution of Households by Cohort (3)	Mean Income by Cohort (4)	Actual 1965 (5)
I. Lorenz-Gini	.4043	4024	.3936	.3994	.3885
II Age-Gini	.2344	.2344	.2164	.2258	.2073
III. Paglin-Gini	1699	.1680	1772	1736	.1812

Sources: Computations by authors from *Survey of Economic Opportunity*, 1966 and the *Current Population Survey*, 1973, based on census money income for all household units classified into 24 groups

(1) Actual 1972: 1972 cohort means, 1972 cohort distribution, 1972 cohort-specific Gini coefficients.
(2) 1972 cohort means, 1972 cohort distribution, 1965 cohort-specific Gini coefficients.
(3) 1972 cohort means, 1965 cohort distribution, 1972 cohort-specific Gini coefficients.
(4) 1965 cohort means, 1972 cohort distribution, 1972 cohort-specific Gini coefficients.
(5) Actual 1965 1965 cohort means, 1965 cohort distribution, 1965 cohort-specific Gini coefficients.

may not choose to alter, depending on normative judgments. We will refer to this latter portion of inequality as nonfunctional or policy-relevant inequality. Functional inequality in this instance "...reflects society's need for varying income over the life cycle as well as other basic facts relating to productivity, investment in human resources, and the work-leisure preferences of households, but only in an average way insofar as these factors express themselves through the age variable" (Paglin, p. 602).[1]

II

Paglin implies that the Paglin-Gini measures only that portion of inequality due to variations in income within cohorts. He states that the change in the Paglin-Gini between 1947 and 1972 "reveals the decline in interfamily inequality of income obscured by changes in the age-income profile and in the age composition of the population" (p. 605). However, as N. Bhattacharya and B. Mahalanobis earlier demonstrated, the Paglin-Gini depends upon 1) the distribution of households by cohort, 2) the mean income of each cohort (the age-income profile), and 3) inequality within the cohorts. In Table 1, where we represent interfamily inequality by cohort-specific Gini coefficients, we illustrate the Bhattacharya-Mahalanobis result by showing that the Paglin-Gini is sensitive to all three components.

Row I shows the actual Lorenz-Gini for census money income for all households in 1972 (col. 1) and in 1965 (col. 5). By this measure inequality increased from .3885 to .4043 over the period.[2] Columns 2, 3, and 4

[1]Paglin recognizes that others may choose a partition different from his age-related partition. Paglin's approach, it should be noted, is one proposed and then rejected over two decades ago by George Garvy. While Paglin's choice of age as the basis for defining functional or optimal inequality is explicit, he provides no explanation for choosing this variable, although there may be one.

[2]The Gini coefficients in Table 1 differ from those in Paglin's Table 3. Paglin's Ginis are based on a division of families into 6 age cohorts. We classify all household units into 24 mutually exclusive cohorts by type of living unit (family or unrelated individual) and sex of head, in addition to the 6 age classes used by Paglin. A similar analysis on only family units (Paglin's concept of recipient unit) classified into 20 age classes was also performed. The substantive findings were unaffected

in row I are a series of standardized Lorenz-Ginis, each of which holds constant one of the age-related sources of inequality: the mean income by cohort in column 4; the distribution of households by cohort in column 3; and the cohort-specific Ginis in column 2.[3] Thus, the Lorenz-Gini in column 4 (.3994) reveals what inequality would have been in 1972 if there were no change in the mean income of cohorts from 1965 to 1972, while the other two determinants changed over time as observed. Because the actual 1972 Gini (.4043) exceeds the column 4 Gini, it follows that the change in cohort mean incomes raised Lorenz-inequality over the period. Indeed, because the actual 1972 Gini exceeds each of the values in columns 2, 3, and 4, each of the three components contributed to growing Lorenz-inequality over the period.

Rows II and III present the age- and Paglin-Ginis for 1965 and 1972, and the same standardizations. The 1965 and 1972 values in row II (cols. 5 and 1) show that the actual age-Gini increased from .2073 to .2344. Since the age-Gini does not account for intracohort inequality, the remaining two sources of inequality must have produced that result. The age-Ginis in columns 3 and 4 of row II are both less than the actual in column 1 verifying that changes in both the distribution of households among cohorts and in the mean incomes across cohorts contributed to the increase in the actual age-Gini.

The Paglin-Gini is defined as the difference between the Lorenz- and age-Ginis and is shown in row III. Columns 1 and 5 of that row indicate that Paglin-inequality fell from .1812 to .1699. Because of this change, Paglin claims that nonfunctional inequality —which he associates with income differences *within* cohorts—has decreased. However, as the values in columns 2, 3, and 4 indicate, that is not the case. Because the values in columns 3 and 4 are above the 1972 Paglin-Gini value of .1699, both of these sources—the distribution of house-

[3]Gini coefficients are insensitive to proportional transformations. multiplying each individual's income by a constant leaves the Gini coefficient unaffected We make use of this fact to construct Table 1 To derive column 4, we assign each household to its cohort and calculate the mean income of each cohort for 1965 and 1972. Then, we multiply each household's income in 1972 by the ratio of the mean income of its cohort in 1965 to the mean in 1972. Since the incomes of all households within a cohort have been multiplied by a constant, the cohort-specific Gini coefficients are as in 1972 The number of households in each cohort is also as in 1972. However, since incomes across cohorts have been multiplied by the relevant different amounts, mean incomes by cohort are as in 1965.

Column 3 is derived by first counting the number of households in each cohort in 1965 and in 1972. To hold mean incomes by cohort and cohort-specific Gini coefficients at their 1972 levels while replicating the 1965 distribution of households among cohorts requires two steps First, we gave each 1972 household a weight which is the ratio of the number of households in its cohort in 1965 to the number in its cohort in 1972 Second, we multiplied the 1972 aggregate income of each income class within a cohort by the same ratio Since the number of households in every cohort and their aggregate incomes are multiplied by the same constant, cohort means and Gini coefficients are at their 1972 magnitudes. However, the distribution of households across cohorts is as in 1965.

Column 2 requires three steps combining the standardizations of columns 3 and 4. To hold the cohort-specific Gini coefficients at their 1965 magnitudes while maintaining the 1972 distribution of households among cohorts and the 1972 pattern of mean incomes by cohort we begin, in this instance, with the 1965 number of households in each cohort. Then, we multiply each household's income in 1965 by the ratio of the mean income of its cohort in 1972 to the mean in 1965 (the reciprocal of the ratio used to obtain col. 4) Each household within a cohort and the aggregate income of each income class in that cohort are then weighted by the ratio of the number of households in the cohort in 1972 to the number in the cohort in 1965 (the reciprocal of the ratio used to obtain col. 3). Since the incomes of each cohort are multiplied by the same constant (a product of two ratios) and the number of households in each cohort has been multiplied by a constant, the cohort-specific Gini coefficients remain at their 1965 magnitudes. These transformations yield the 1972 pattern of mean incomes by cohort and the 1972 distribution of households among cohorts. Note that this is the information required to compute Paglin's age-Gini. The age-Gini in column 2 which results from the transformed 1965 data is identical to the age-Gini in column 1 computed from the actual 1972 data as the age-Gini does not reflect intracohort inequality. The Lorenz-Gini coefficients differ as this source of inequality is reflected in the measure.

While these standardizations are illustrative they do not partition the difference between the 1965 and 1972 Lorenz-Gini coefficients into additive terms which just exhaust the total. For a discussion of the difficulties associated with decomposing Gini coefficients, see Graham Pyatt.

holds by cohort and the mean income of cohorts—contributed to a reduction in Paglin inequality. As column 2 indicates, however, the change in the cohort-specific Gini coefficients led to an *increase* in Paglin-inequality between 1965-72. This, then, is the basis of Paglin's misperception of his own measure. Like the Lorenz-Gini, changes in the Paglin-Gini reflect the combined impact of all three sources of inequality, and not simply changes in within-cohort inequality, as Paglin suggests.[4] While all three sources contributed to the increase in Lorenz-inequality from 1965 to 1972, two of the sources operated to decrease the Paglin-Gini. Ironically, only the changes in the cohort-specific Gini coefficients contributed to the increase in Paglin-inequality over this period.

III

Apart from the difficulty in interpreting changes in inequality with the Paglin-Gini, there is the normative question raised by Paglin's procedure of establishing each period's partition between functional and nonfunctional (or policy-relevant) inequality from that period's observed age-income profile. Presuming that some partitioning of inequality based on life cycle considerations is meaningful, it does not follow that all changes in any of the determinants of the age-income profile should be exempt from a policy judgment or a policy response, as Paglin's procedure implies.[5]

This point can be illustrated with a simple heuristic model. Assume that the set of underlying determinants of the measured *P*-reference line in any year is exhausted by:

- *A*: The distribution of inherent physical and mental capabilities by age;
- *B*: The returns to "learning-by-doing" (experience) by age;
- *C*: The returns to investments in human capital by age;
- *D*: Income transfers by age;
- *E*: Returns from physical capital by age;
- *F*: Earnings effects of labor-leisure choices by age;
- *G*: Distribution of families by age.

In Figure 1, the likely effect on the age-income profile of determinants *A F* is depicted.[6] Holding tastes for work among age cohorts (at, say, the average for the population) constant, inherent capabilities *A* would likely give the age-income profile a slight peak, because of declining capabilities after some age. The addition of work experience *B* would tend to increase income with age until the marginal return to experience falls to zero. Similarly, the returns to education *C* would add income at all ages, but especially young and middle ages, again increasing the peakedness of the profile. To the extent that public income transfers *D* and returns from nonhuman capital *E* favor the old relative to the young and middle-aged cohorts, the income line would continue to rise throughout but would be less peaked. Finally, adding variations in hours worked due to leisure-income preferences *F* would lead to a decrease in expected income for aged cohorts and increases for the young and middle aged.[7] This determinant contributes substantially to the peakedness of the age-income profile. The cumulative effect of all these variables is the observed age-income profile—the heavy

[4]Paglin is correct in asserting that the age-Gini is insensitive to changes in within-cohort inequality as shown by row II, columns 1 and 2.

[5]Acknowledging the weaknesses of the Lorenz-Gini as a standard of equality since it does not allow for life cycle income variation, it would seem more appropriate to confront the life cycle needs issue by a collectively established norm reflecting a social judgment on the constitution of life cycle needs. Such a norm would be immune from transitory changes in the myriad of variables underlying any year's observed age-income profile. Reflecting a collective judgment on age-related needs, this norm would be analagous to the official *U.S.* poverty definition, and like that definition, would change only after explicit deliberation.

[6]Determinant *G* does not influence the age income profile. However, because it is used to weight observations on that profile in calculating the *P*-reference line, it is reflected in the age-Gini.

[7]The oversimplicity of this scheme becomes obvious at this point. Labor-leisure choices depend not only on tastes but also on the incentives created by the other determinants, notably *D*. A more complete model would account for these interdependencies.

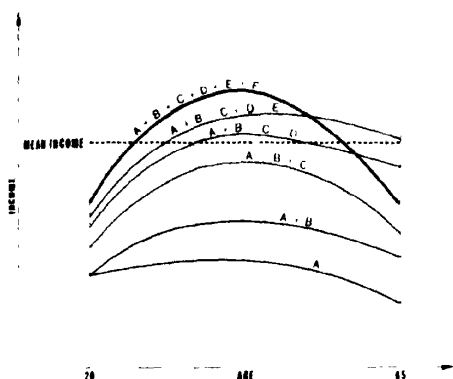



FIGURE 1. COMPONENTS OF THE AGE-INCOME PROFILE

line $A + B + C + D + E + F$. The more peaked the age-income profile the higher the age-Gini, *ceteris paribus*. However, the P-reference line derived from the age-income profile and the age-Gini which is calculated from it reflects one additional variable: the distribution of units by age, G. The heavier the concentration of units in cohorts with income far from the average income over all ages (represented by the dotted heavy horizontal line), the greater the age-Gini.

This, then, is the sort of framework which underlies the derivation of the P-reference line and Paglin's partitioning of inequality into that which is functional and that on which normative policy judgments must focus. Because his partitioning is an annual exercise, any change in variables A F leading to increased peakedness in the age-income profile, or any change in the age distribution toward cohorts whose expected income is above or below the mean, will cause an increase in functional inequality (the age-Gini), an increase in the Lorenz-Gini, and an indeterminate effect on the Paglin-Gini.

To illustrate the problems posed by such an approach for evaluating the historical record, return to Table 1. As noted, because each of the Ginis in columns 2, 3, and 4 of row I is less than .4043, each of the three determinants of the Lorenz-Gini contributed to the increase in inequality over the period. Row III, however, tells a conflicting story. For example, column 4 shows that if the 1965 cohort mean incomes had persisted through 1972, the Paglin-Gini would have been .1736 rather than the observed .1699—implying that the observed change in the pattern of mean incomes by cohort reduced inequality as measured by the Paglin-Gini. A similar contradiction holds for the change in the distribution of households among cohorts (col. 3). These contradictions arise because a change in any of the three components alters *both* the Lorenz-Gini and the age-Gini, but alters them by different amounts. In this particular case, changes in the pattern of mean incomes by cohort alters the age-Gini by a greater absolute amount than the Lorenz-Gini. Hence, Paglin's preference for the scenario of row III relative to that of row I requires that he accept this larger impact of the changing pattern of cohort mean incomes as reflecting a functional source of growing inequality not reflected when the same effect is simply standardized out of the Lorenz-Gini (as in row I).

To further illustrate this problem in Paglin's measure, suppose that a transitory change in the pattern of returns to experience had been responsible for the increased peakedness in the pattern of mean incomes among the age cohorts. In particular, assume that it resulted from a rise in the minimum wage, a consequence of which is that large numbers of young workers face intermittent unemployment. Assume further that the increased steepness of the profile raises the age-Gini more than the Lorenz-Gini. Does Paglin really wish to claim that in this circumstance the degree of inequality with which policy makers should be concerned has been reduced? Paglin's method requires an affirmative answer to this question. The burden of demonstrating why this should be so falls on him.

IV

Another issue raised by Paglin's paper concerns the use of inequality measures for evaluating the effectiveness of public income transfer programs. This matter has

both normative and empirical aspects, which we will consider in turn.

Implicit in Paglin's framework is a criterion for judging the effectiveness of income transfers, if the objective is to reduce inequality. An income transfer is "Paglin-efficient" only if it reduces the variation of incomes within an age cohort; transfers which involve intercohort redistribution are by definition "Paglin-inefficient." Thus, given a fixed value of income to be redistributed, the largest decrease in the Paglin-Gini can be obtained only through strictly intracohort transfers.

A criterion for judging the effectiveness of income transfers is also implicit in the Lorenz-Gini. Any transfer from a household with high current income to one with low current income is Lorenz-efficient without regard to the age of the household head. To the extent that the current income of the young and the old are downward-biased measures of economic welfare, Lorenz-efficient transfers may be judged ineffective. While Paglin's measure avoids this problem, it creates another. Transfer programs which are widely supported—for example, the Social Security system—turn out to be Paglin-inefficient.

The difference between Lorenz-efficient and Paglin-efficient transfers is illustrated in Table 2. The first panel depicts a hypothetical economy of 6 units divided equally between two age cohorts. The young cohort has a mean pretransfer income of $150; the old cohort a mean of $50. The second panel shows the economy after $150 of Paglin-efficient transfers have been made. (The numbers in parentheses are the positive and negative transfers.) By undertaking these transfers, complete Paglin-equality is achieved with all individuals within an age cohort having the mean income of that cohort. Notice that in this Paglin-efficient transfer program, the young individual with $150 of pretransfer income pays no taxes, while the old individual with but $90 of income pays $40 in taxes. The income range has fallen from $250 to $100. In the third panel, a more conventional tax-transfer system designed to eliminate both high and low income extremes, regardless of age, is depicted. Again, the total transfer budget is $150. The highest taxes are imposed on those with the most income; the poorest receive the largest transfers. Through this scheme, the mean income of the young cohort has been reduced and that for the old cohort has been raised. Indeed, Paglin-equality has been achieved for the old cohort, and the range of incomes has been reduced from $250 to $70. By Paglin's criterion, the transfer system represented in the second panel is efficient; that in the third panel is not. However, the transfer system in the third panel achieves a greater reduction in Lorenz-inequality than that in the second panel. More importantly, it does so without taxing away more income from poorer individuals than from those with higher incomes. Which of these two outcomes is preferred depends on the precise meaning of "young" and "old." Where young and old refer to age cohorts for which current income is a satisfactory proxy for economic welfare, the Lorenz standard would be preferred; for those cohorts in which current income deviates from eco-

TABLE 2—ALTERNATIVE TRANSFER SYSTEMS

	Individuals 1	2	3	Age Cohort Mean Income
Pretransfer Income				
Young	$40	$150	$260	$150
Old	10	50	90	50
Posttransfer Income (after $150 of Paglin-efficient transfers)				
Young	150	150	150	150
	(+110)	(0)	(−110)	
Old	50	50	50	50
	(+40)	(0)	(−40)	
Posttransfer Income (after $150 of conventional transfers)				
Young	90	120	150	120
	(+50)	(−30)	(−110)	
Old	80	80	80	80
	(+70)	(+30)	(−10)	

nomic welfare, Paglin's criterion might be appropriate.[8]

In this context, it should be noted that, as calculated, the age- and Paglin-Gini coefficients incorporate transfers which are by definition Paglin-inefficient, since Paglin's income concept is census money income.[9] For example, in 1972, mean pretransfer income of aged male-headed families was $5337 while mean census money income was $8372. Public cash transfers have grown from a few billion dollars in 1947 (the first year of Paglin's time-series analysis) to a level 20 times that amount in the last year of his analysis.[10] Because the bulk of this growth has been in age-related Social Security benefits, the shape of the age-income profile over time is significantly affected. Consequently, Paglin's income profiles are based on an inappropriate definition of income which biases his conclusions on the trend of functional inequality in the postwar period.[11]

There is a further point. Paglin accepts the different age-income profile in census money income in each year as socially optimal. Because public transfers are included in this income definition, Paglin has rejected the optimality of the age-income profile generated by the market. Yet no explanation is given as to why this particular combination of the market and political decisions should yield the optimal profile.[12] If one is to rely on the observed profile as a norm, a more consistent position would be that either the market alone or the market plus all political decisions would yield a superior measurement. Apparently it is the ready availability of published data which has dictated Paglin's choice of income concept.[13]

[8]Again, it should be emphasized that any transfer which alters the age-income profile will change both the Lorenz-Gini and the age-Gini. Consider, for example, a transfer scheme which taxed the middle-aged rich and transferred to the young and old poor, such that no person taxed had less aftertax income than any person benefitted had after the transfer. Clearly, such a scheme reduces Lorenz-inequality. Yet, in Paglin's approach, the transfer would reduce the peakedness of the age-income profile, and shift both the P-reference curve and the Lorenz curve toward the 45 degree line. While the Lorenz-Gini and the age-Gini would unambiguously decrease, the Paglin-Gini might rise, fall, or remain constant.

[9]Census money income includes wages and salaries, dividends, interest, rents, and cash transfers before taxes, but excludes capital gains, in-kind transfers, the services of owner-occupied housing, and benefits of publicly provided goods and services.

[10]According to the *Current Population Survey (CPS)*, public cash transfers totalled $80 billion in 1972, more than 10 percent of market generated income.

[11]Were income measured properly there would still be a problem. Paglin's age-income profile is derived from an implicit regression model in which income is taken to be dependent on age ***and age alone***. There are several reasons for doubting that such a regression would yield an unbiased estimate of the determinants of functional inequality as defined by Paglin. On the one hand, age is not a good proxy for important market factors which determine the age-income profile—work experience, the value of benefits from investing in human capital, and tastes for income vs. leisure. On the other hand, age is collinear with race, urban-rural location, education, size of city of residence, and other variables which Paglin would presumably wish to include in his definition of nonfunctional or policy-relevant inequality. Thus, if age were the only independent variable the estimated profile would be biased. An indication of the bias can be obtained by regressing family income first on only age and (age)^2 and then on these two variables plus other economic and demographic variables and comparing the coefficients on the age variables. When such regressions were estimated with microdata from the 1972 *CPS*, the coefficient on the age variables changed substantially. The estimated age-income profile was significantly more peaked in the first regression. The coefficients were 855.6 and −9.3, respectively, and the age-income profile was: age 20, $13,392; age 50, $19,533; and age 60, $17,861. In the more comprehensive regression, we added dummy variables for region (3), city-type (3), race of head (1), and sex of head (1); while family size, education and (education)^2 entered continuously. The age and (age)^2 coefficients change to 657.4 and −6.2, respectively, and the age-income profile becomes age 20, $10,682; age 50, $17,454; and age 60, $17,145.

[12]In the age-income profile used by Paglin, nonwage income, influenced in part by inherited wealth, is also included. While Paglin presumably would not wish to include income differences stemming from inherited wealth differences in his category of functional inequality, to the extent that nonwage income from inherited wealth is not age-neutral, his empirical procedure does just that.

[13]Alternatively Paglin could employ the age-Gini of pretransfer income in deriving the P-reference curve and then subtract it from the posttransfer Lorenz-Gini. This two-step comparison, however, could not

V

For both normative and computational reasons, then, we are led to reject the new measure of inequality proposed by Paglin, and with it the reliability of his conclusion that "estimates of inequality have been overstated by 50 percent" (p. 608), and that since 1946 there has been a considerable reduction in net inequality. We have argued that his decision to base a definition of nonfunctional inequality on annual age patterns of posttransfer income (as opposed to pretransfer or "full income") is arbitrary and without a rationale, leads to biased estimates of the trend of postwar inequality, and limits the usefulness of his measure as a gauge of the redistributive effect of income transfer policies.

Paglin has provided a new measure of inequality which is not unique, since "Once we cast aside the socially unrealistic 45 degree line of equality, we are free to generate new reference lines corresponding to explicit and reasonable definitions of equality, equity, or Pareto optimality" (p. 599). An inequality measure which allows for life cycle variations is appealing. However, such a standard requires an explicit judgment on the optimum life cycle pattern, and relying on annual observations of an arbitrarily observed pattern is unsatisfactory. Indeed as we have tried to make clear, the normative underpinning of the Paglin-Gini would be at odds with conventional notions of equity, even if it were successfully measured.

be calculated from published data and a major advantage of the Paglin measure would be lost. Furthermore, Garvy demonstrated that the subtraction of Gini coefficients from separate underlying distributions is uninterpretable

REFERENCES

N. Bhattacharya and B. Mahalanobis, "Regional Disparities in Household Consumption in India," *J. Amer. Statist. Assn.*, Mar. 1967, *62*, 143-61.

G. Garvy, "Inequality of Income: Causes and Measurement," in *Eight Papers on Size Distribution of Income*, Nat. Bur. Econ Res., *Stud. in Income and Wealth*, Vol. 15, Princeton 1952.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598-609.

G. Pyatt, "On the Interpretation and Disaggregation of Gini Coefficients," *Econ. J.*, June 1976, *86*, 243-55.

U.S. Bureau of the Census, *Current Population Survey*, Mar. 1973 (data tape).

U.S. Office of Economic Opportunity, "Survey of Economic Opportunity," (*SEO*) conducted spring 1967, available on tape Data Bank, Univ. Wisconsin.

The Measurement and Trend of Inequality: Comment

By JOSEPH J. MINARIK*

Morton Paglin's recent article in this *Review* is an important contribution to the analysis of the distribution of income. He argues convincingly that inequality of incomes on a life cycle basis would exist even in a perfectly equalitarian society. He then provides a clear and simple technique by which inequality in excess of that related to the age-income profile can be measured. Paglin finds that on this basis the degree of inequality in the distribution of income fell significantly over the quarter century from 1947 to 1972; that is, the difference between the actual distribution and that which would have obtained if each family received the mean income for its age group decreased.

Paglin's conceptual approach is a meaningful improvement over consideration of Gini coefficients without reference to underlying population change. At the same time, the Paglin technique raises further questions which shadow his conclusions on the trend of inequality. The objective of this comment is to discuss some of these questions and suggest alternative applications of the technique; the results will indicate that Paglin's technique must be used with caution for conclusions regarding the trend of inequality to be fairly drawn.

The remainder of the comment will be divided into three sections. The first will consider Paglin's selection of the age-income profile as a measure of "permissible" income inequality. The second will examine the distribution of total family income as an indicator of equality. The third section will be a brief conclusion.

I. How Much Inequality is Inevitable?

Paglin's central point is that perfect equality is an unrealistic goal for incomes measured over a one-year period. Most economists agree that greater equality of lifetime incomes is the real objective (abstracting from long-term growth of real incomes) and that inequality on a short-term basis is inevitable and tolerable. Paglin identifies the inequality of incomes by age as a basic source of inevitable inequality in annual incomes, because younger and older workers are not as productive as workers of prime age. He then reasons that the least inequality we could expect would be perfect equality along the age-income profile (that is, all workers of the same age have the same income, while incomes of young and old workers are below the mean). The Paglin-Gini is the difference between the Gini coefficient for the actual distribution and the Gini coefficient which would have obtained had there been perfect equality along the age-income profile, the age-Gini. Paglin chooses the family as his unit of observation, omitting unrelated individuals, and uses total family income as his income measure. He computes Paglin-Gini, age-Gini, and Lorenz-Gini coefficients using data for 1947 through 1972 and finds a falling trend of the difference between each year's age-Gini and Lorenz-Gini.

Paglin assigns the increase in the age-Gini (the degree of inequality of the age-income profile) to "... the expansion of higher education ... and to the increase in the percent of the aged and young adults in the population" (p. 604). His technique measures the excess of inequality over that caused by the trend toward a less favorable

*Research associate, Economic Studies Program, The Brookings Institution. Henry J. Aaron, John L Palmer, and Joseph A. Pechman made helpful suggestions but should not be implicated in any errors. The research on which this comment is based was supported by a grant from the National Science Foundation. The opinions expressed herein are mine and should not be attributed to the officers, trustees, or staff members of the Brookings Institution, or to the National Science Foundation.

age-income profile in the Paglin-Gini, but it does not explicitly allow for the inequality caused by increased investment in higher education.[1] In a perfectly equalitarian society individuals could still have a free choice to postpone their working years and invest in further schooling, and thereby earn a fair return on their capital and opportunity costs in the future. It would seem to be a reasonable exercise to relax Paglin's implicit assumption that all families with heads of the same age should have the same income regardless of the schooling of the head, and exclude from the area of excess inequality that caused by increased human investment. Indeed, one could include other family attributes as factors leading to "inevitable" inequality -- the sex of the family head, the number of earners, even the color of skin. Which factors should be considered, and whether they should be considered for diagnostic purposes ("All things considered, what has been happening to the income distribution?") or for policy-making purposes ("What should we do, if anything, about the income distribution?") must be determined according to personal values as well as scientific principles. Calculations will be presented here for a simple and relatively uncontroversial variant of Paglin's technique: measuring inequality around separate age-income profiles for classes of families whose heads have similar schooling.

These calculations will be made here through an age-schooling-Gini coefficient, which will be calculated from the income distribution which would obtain if all families had the mean income of the class of families with their head of the same age and schooling.[2] Testing the joint effects of age and education on income inequality requires the use of a microdata set to provide income cross tabulations. The broadest time span of data sets similar to Paglin's source is the *Current Population Survey (CPS)*; the 1967 and 1971 through 1974 income years were available to the present author. The *CPS* is the survey on which the aggregate data used by Paglin are based, but it is not available as a microdata set for income years before 1967.

Columns (2) through (4) in Table 1 are recalculations of Paglin's Paglin-Gini, age-Gini, and Lorenz-Gini coefficients using the available *Current Population Survey* files. The income distribution is almost constant in 1967 and 1974 while Paglin's Paglin-Gini falls, showing an unfavorable shift in the age-income profile. Paglin argues that the distribution has become more equal apart from this shift.

Columns (4) and (5) in Table 1 show the age-schooling-Gini and the adjusted Paglin-Gini coefficients. These figures show a distinctly different trend from the Paglin-Gini and age-Gini.[3,4] The age-schooling-Gini

[1]I have noted the trend toward a less favorable age-income profile over time and provided quantitative estimates of its effect in my doctoral dissertation. My conclusions on the distributional effect of higher education are diametrically opposed to those of Paglin, as will be evident shortly.

[2]The age categories used here are the same as Paglin's: 14-24, 25-34, 35-44, 45-54, 55-64, and 65 and over. The schooling categories are incomplete elementary school, complete elementary school, complete elementary school plus some high school, complete high school, complete high school plus come college, and complete college

[3]The adjusted Paglin-Gini is lower than the Paglin-Gini simply because it includes another dimension of inequality, that on the basis of schooling, in the area of "inevitable" inequality. The level of these coefficients is obvious and uninteresting, what is of interest is their relative trends

[4]The Gini coefficients used here were calculated by the formula

$$G = \sum_{i=2}^{n} \left[\left(\frac{\sum_{j=1}^{i-1} N_j}{\sum_{i=1}^{n} N_i} \right) \left(\frac{\sum_{k=1}^{i} T_k}{\sum_{i=1}^{n} T_i} \right) - \left(\frac{\sum_{k=1}^{i} N_k}{\sum_{i=1}^{n} N_i} \right) \left(\frac{\sum_{j=1}^{i-1} T_j}{\sum_{i=1}^{n} T_i} \right) \right]$$

where G is the Gini coefficient, N_i is the number of cases in the ith class, T_i is the total income in the ith class, and there are $i = 1, n$ classes ranked by income. The income classes were hundred dollar intervals from -\$9,900 to \$50,000. No spline function such as Paglin's was used because most of the present exercises utilized a large number of cells and a spline function would not affect the measured trend of inequality. Paglin's Lorenz-Gini coefficients are .355, .356, and .359 for 1967, 1971, and 1972, respectively. The figures here are within 3 percent of Paglin's; the reason for the understatement is the truncation of incomes above \$50,000 in the Public Use Samples for purposes of

TABLE 1—MEASURES OF INEQUALITY OF DISTRIBUTION OF TOTAL INCOME AMONG FAMILIES; SELECTED YEARS, 1967-72
(Figures are rounded)

Year (1)	Lorenz-Gini (2)	Age-Gini (3)	Paglin-Gini (4)	Age-Schooling-Gini (5)	Adjusted Paglin-Gini (6)
1967	.346	.107	.238	.180	.166
1971	.351	.115	.236	.185	.166
1972	.354	.116	.239	181	.173
1973	.352	.115	.237	.181	.171
1974	349	.116	.233	.180	.169

Sources: Derived from *Current Population Survey* March Income Supplement Public Use Sample data tapes of 1968 and 1972-75.

TABLE 2—MEASURES OF INEQUALITY OF DISTRIBUTION OF EARNED INCOME AMONG FAMILIES; SELECTED YEARS, 1967-72
(Figures are rounded)

Year (1)	Lorenz-Gini (2)	Age-Gini (3)	Paglin-Gini (4)	Age-Schooling-Gini (5)	Adjusted Paglin-Gini (6)
1967	.396	.144	.253	.216	.179
1971	.415	.161	.254	.231	.185
1972	.421	.163	.258	.227	.193
1973	.423	.167	.256	.231	.192
1974	.427	.170	.258	.235	.192

Sources See Table 1.

hardly changes over the period, resulting in an increase in the adjusted Paglin-Gini. So while the Paglin-Gini, using the age-income profile for a base, finds a 2 percent *decrease* in inequality, the adjusted Paglin-Gini, based on separate age-income profiles for groups with different schooling attainments, finds a 2 percent *increase* in inequality.

II. What Kind Of Income Should Be Measured?

Paglin's measurements and all those thus far in this comment have been made on total family money income. Total income is an acceptable proxy measure of potential consumption within the year, but we might also want public policies to equalize the distribution of earned income. Paglin uses his results to refute the claim that "... there has been no significant reduction of inequality from 1947 to 1972 despite the massive spending on education and training programs, the more generous cash and merit good transfers, and the legislative and judicial actions directed at bringing minorities and underprivileged groups into the mainstream of the economy" (pp. 603-04). While Paglin's results may say something about potential consumption, they do not bear at all on the bulk of the rhetoric in this statement, which suggests equality in earning power. Accordingly the earlier calculations in this comment were duplicated for earned income, including wages, salaries, and farm and nonfarm self-employment income as defined by the Census Bureau.

Table 2, which shows these recalculations, is a much less attractive picture to the equalitarian than Table 1. All of the measures—the Lorenz-Gini, the Paglin-Gini

confidentiality. Paglin's Paglin-Gini coefficients are .245, .237, and .239, respectively; again my figures are well within 3 percent tolerance Paglin's Paglin-Gini coefficients could be used in the textual analysis without in any way affecting the conclusions. Column (4) equals column (2) minus column (3); column (6) equals column (2) minus column (5).

and adjusted Paglin-Gini—show increasing inequality, the Lorenz-Gini by 8 percent, the Paglin-Gini by 2 percent, and the adjusted Paglin-Gini by 7 percent.[5]

Obviously the results in both Section I and Section II of this comment are strictly limited in coverage. The time period, constrained by data availability, is quite short; macroeconomic conditions shroud any long-term trend information available.[6] The point, however, is simple, and it stands clear of the limitations: measurements of inequality depend totally on what you measure and how you measure it, and Paglin's technique is no exception.

III. Conclusions

It is necessary in closing to restate an appreciation of Paglin's contribution to the study of income distribution. He has provided a simple and eminently useful tool for distribution analysis.

This comment made two major criticisms: 1) Paglin's age-Gini, or measure of inevitable income inequality in a short-term period, could well include population attributes beyond the simple age-profile of income. One simple addition made here, the use of separate age-income profiles for different schooling classes, completely reversed Paglin's trend in inequality over an admittedly but necessarily short period. 2) In measuring our progress towards a truly "equalitarian" society, it was suggested that the distribution of earned income might be a more appropriate indicator than that of total income. A duplication of the earlier calculations on total income revealed a far more pessimistic picture than Paglin's. The conclusion is that both a baseline inequality measure and an income measure must be explicitly chosen before conclusions can be drawn from Paglin's technique.

There are two empirical factors at work in these results which will be important in future work on income distribution. One is the dramatic shift in relative rewards of highly educated workers. The recent low returns of young, highly educated family heads have much to do with Paglin's sharply peaked age-income profiles and his rising age-Gini coefficients. The falling returns to higher education raise important questions about the meaning of lifetime income equality and the future shape of the income distribution.

The second factor is the growing proportion of families who have no earned income at all. While the distribution of earned income among families with such income has become more equal, this trend is reversed when families with no earned income are taken into account.[7] It is this factor which challenges the record of public policy and Paglin's assertion of a trend toward greater equality.

[5]These results are not affected by excluding households with head 65 and older from the calculations, or by changing the income concept to total income less public transfers. The results would be affected by excluding families with no earned income from the calculations; this will be discussed briefly in the conclusion.

[6]Linear regressions indicate that a higher unemployment rate significantly increases the Paglin-Gini.

[7]See my dissertation, chapters 1 and 8.

REFERENCES

J. J. Minarik, "A Microanalysis of the Size Distribution of Income," unpublished doctoral dissertation, Yale Univ. 1974.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer Econ. Rev.*, Sept., 1975, *65*, 598–609.

U.S. Bureau of the Census, *Current Population Survey*, Mar. Income Suppl. Public Use Sample data tapes, 1968; 1972–75.

The Measurement and Trend of Inequality: Comment

By C. JOHN KURIEN*

In a recent article in this *Review*, Morton Paglin argues that his "more basic" measure of interfamily income inequality, which he calls the Paglin-Gini, obtained by subtracting the age-Gini from the Lorenz-Gini enables him to capture true variations in income by removing the effects of variations in income due to age. This note comments on some theoretical and statistical issues underlying Paglin's method and interpretation, clarifying the meaning of what he has attempted to do and some of the limitations of his method and interpretation.

As Paglin recognizes (fn. 3, p. 601), Gini coefficients are monotonically related to standard deviations. For a given population size n, the Gini is then monotonically related to the total variation in income $n\sigma^2$. Following traditional analysis of variance procedures, this total variation may be *partitioned* into components in various ways, one such partitioning being between age related variation $n\sigma_a^2$ and its residual or "Paglin-variation," $n\sigma_p^2 = n\sigma^2 - n\sigma_a^2$. If Gini coefficients are constructed for these partitioned variations, we get the age-Gini and the Paglin-Gini. The procedure is also equivalent to calculating the standard error of estimate from a regression of age on income using a suitable non-linear function (as Paglin uses) and associating the standard error of estimate to Paglin-Gini while the standard deviation of incomes is associated with Lorenz-Gini. Age-Gini is associated with the variability "explained" by age. Therefore, it is clear that the usual caution given to interpretations in analysis of variance in regressions should be given to interpretations based on Paglin's method also. Specifically, when Paglin identifies the Paglin-Gini as a truer measure of interfamily income inequality than the Lorenz-Gini, his statements need to be heavily qualified.

In order to clarify the distinction between what he attempts to do and what he succeeds in doing, a very brief theoretical discussion is useful as a starting point. Measures of inequality are of interest primarily as aids to measuring interpersonal differences in welfare. Income is conceived as the best available proxy[1] for welfare level or "utility" level. Then, if for each individual there is a time distribution of incomes such that incomes vary over the lifetime of the individual, two individuals who differ with respect to no characteristic but age may have different incomes. This difference in income reflects only difference in the stage of the life cycle in which the two individuals are, and not any true differences in welfare between them. If such fluctuations are identified as life cycle fluctuations in income, eliminating these fluctuations improves income distributions as instruments for interpersonal comparisons of welfare. The elimination of these "spurious" differences is implicitly associated with the elimination of age-related variation in Paglin's study. The appeal of Paglin's study, perhaps, is related to this association.

However, in the Paglin-Gini, he not only eliminated a large part of such spurious life cycle adjustments, but some true differences in welfare as well. Thus, though he recognizes (pp. 601–02) that in a society with steadily growing incomes, the older generations are on the average poorer than younger generations he ignores the fact that elimination of this difference which is also captured by the age-Gini eliminated some

*Associate professor of economics, McGill University. I am obliged to C. Green and P. Davenport for comments and suggestions.

[1]A good case can be made for consumption (not consumer expenditure) as an even better proxy, but measurement problems in this case are even more difficult.

true inequalities in welfare as well.[2] To the extent that age-related variations include more than pure life cycle adjustments, the Paglin-Gini understates true income inequality.

The intergenerational inequalities in welfare may be due also to factors other than steady growth in productivity. Individuals whose earnings were disrupted by the depression and World War II during the years in which their earnings would have been high are genuinely poorer over their lifetime than those who did not so suffer.[3] If life expectancy goes up to a degree more than anticipated during most of an individual's life, in later years persons so affected may be genuinely poorer. Changes in family structure from one in which the children assumed a greater responsibility for the well-being of their elder relatives to one in which they do not, could reduce the old to lower levels of real income than anticipated. If opportunities for participation in the labor force for females improved over time, then single young females and the young families may genuinely be better off. The important point is that the cross-sectional age-income profile captures some true interfamily differences in welfare in addition to some intrafamily differences and thus eliminating the age-Gini eliminates much more than life cycle adjustments.

Further, all life cycle adjustments in welfare are not captured by the age-Gini. If two individuals A and B of the same age faced with identical opportunities choose different occupations with identical lifetime discounted total incomes, but with different age-income profiles, in any one year they may have different incomes. The age-Gini does not capture these differences and hence eliminates too little as well.[4]

It should also be pointed out that there are several other factors similar to age which contribute to similar spurious variations. Because of differences in living costs in different regions, part of interregional differences in mean incomes is spurious, though some such variation is real. Differences in family incomes reflect, in part, the degree of participation in the labor force (for example, how many adults in the family shall work outside the home?) and in part differences in opportunities available to the families. Similarly incomes also differ due to differences in the number of hours worked in a year; part of these differences reflect opportunities available, but part reflect only choices related to differences in tastes. Thus it is important to recognize that we can partition total variation in incomes in various ways; between age-related and residual variation as Paglin does, labor-force-participation-rate-related and residual, region-of-work-related and its residual, and so on. None of these partitions, unfortunately, reflects entirely differences in choices, all reflect differences in opportunities as well.[5] From a welfare point of view, perhaps, the most useful partition is *variation related to differences in opportunities among persons* and its residual, *choice-related variation*. The former measures differences in welfare among individuals or true income inequality while the latter is spurious for the purpose of interpersonal comparisons of welfare. Thus, an ideal measure of income distribution will eliminate all choice-related variation, but none of differential-opportunity-related variation.[6]

Paglin's effort may, then, be viewed as follows. He attempts to remove a part, but only a part, of choice-related variation in incomes. In the process, he eliminates some

[2]While Paglin is correct in asserting that intergenerational differences are almost impossible to eliminate, it does not follow that we accept these differences as they are; societies can and do make efforts to reduce these inequalities through public policy measures, implicity recognizing that they are real.

[3]Societies sometimes compensate those who were affected by wars by veterans' benefits.

[4]In fairness to Paglin it must be pointed out that this problem arises entirely due to income being an imperfect measure of welfare and no simple adjustment to income distribution can remove this liability.

[5]Each such partition may be further partitioned into differential opportunity-related variation and a choice-related variation. Therefore, eliminating any such variation eliminates too much.

[6]It may be noted here that we use family incomes rather than individual incomes in an effort to remove what appears to be largely, but not entirely, spurious *differences in incomes among members of the family*. That this too is imperfect is clear from earlier observations about labor force participation choices.

true interpersonal differences in welfare as well. While the effort is in the right direction, its success is somewhat limited. However, the difficulty of proceeding much further in the direction he has chosen is also clear when we recognize that all choices among income and nonincome benefits ("noneconomic benefits") as well as choices among time-income profiles distort measures of welfare inequalities using income distributions.

Moving to Paglin's conclusions, unfortunately, it is difficult to accept the definiteness of his conclusions. Larger changes in the Paglin-Gini than in the Lorenz-Gini definitionally follow from substantial changes in the age-Gini and we have to ask the question whether changes in the age-Gini reflect changes in true (i.e., differential-opportunity-related) inequality or only spurious ones (choice-related). As the proportion of population whose lifetime incomes were disturbed by economic turbulences of 1930's and 1940's declines (up to a point), greater variation in true incomes appears as related to ages of the individuals, and age-Gini should be expected to increase as a consequence. Other factors mentioned earlier also may play a similar role. Therefore much more study is required before we can conclude that changes in age-Gini capture only changes in age structure of the population and life cycle patterns. Without such a conclusion, it is difficult to relate changes in age-Gini to changes in income distribution unrelated to changes in welfare.

On the wealth distribution too, we run into similar problems. If a large part of the inherited wealth is passed on at the death of the legator, then the probability of having inherited wealth increases with age at least up to a point, so that older individuals on the average should be expected to have greater inherited wealth. Some of the wealth differences in the age cross section are, therefore, opportunity-related and Paglin's revision here also appears to be too strong.

The weakest part of his conclusion is related to the revaluation of the share of the lowest quintile. The fact that a high proportion of those in the lowest quintile are very old or very young is true enough, but to assume away the possibility that these groups may be truly poorer is strange indeed. It is almost inevitable that those who have lower average incomes in any partition will be found with greater relative frequency in the lower quintile than in the upper four quintiles, whether the differences are opportunity-related or choice-related. In any case, if a certain degree of inequality and its social manifestations are associated with a certain value of an income inequality measure, what good is it to conclude that these same social situations are related to a smaller number for income inequality in a "better" measure?

REFERENCES

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598–609.

G. U. Yule and M. G. Kendall, *Introduction to the Theory of Statistics*, London 1957.

The Measurement and Trend of Inequality: Reply

By MORTON PAGLIN*

I shall deal with the five papers in sequence, concentrating initially on the main issues raised in each; this will be followed by a review of miscellaneous points. The heart of the controversy involves the validity and meaning of the Paglin (P)-Gini: is it based on an erroneous disaggregation of the Lorenz (L)-Gini or does it correctly measure a socially important dimension of inequality? To answer this I will attempt a microeconomic analysis of the Gini with particular attention to the interaction terms. When this issue is clarified, many other points can be more easily resolved.

Eric Nelson argues that I have used an incorrect statistical procedure in subtracting the age-Gini from the L-Gini; instead, the "true" measure of my P-Gini is "the sum of the product of three terms": $P = \Sigma\lambda_i\gamma_i G_i$ (his equation (4)). He really doesn't quite mean this, because with population shares and income shares each adding to one, the summation of the double-weighted cohort Gini ratio, cut loose from the interaction term, yields an absurdly low estimate of within-cohort inequality (.061 for 1972). Hence in the rest of the paper he uses one or the other set of weights, or suggests dividing the right side (only) of equation (4) by $\Sigma\lambda_i\gamma_i$. Either way he departs from the logic of the disaggregation suggested in his equations (1) through (4). I will return to this point later.

I agree with Nelson that the age-cohort income distributions overlap, and that when I subtract the age-Gini

$$\sum_{i \neq j} p_i p_j \mid \mu_i - \mu_j \mid \cdot \frac{1}{2\mu}$$

from the L-Gini, I am left with $\Sigma\lambda_i\gamma_i G_i$ plus an interaction term. The crucial issue which he fails to discuss is the economic meaning of the interaction effects and the case for including or excluding them from an overall measure of intracohort inequality. They are clearly a major part of the $\Sigma \mid x_i - x_j \mid$ term in the Gini and we must determine on a microeconomic level the significance of these effects before we can decide what to do with them.

N. Bhattacharya and B. Mahalanobis (whom Nelson cites but does not follow) offer an intuitively appealing argument for allocating the interaction term to the within-cohort box. They proceed to partition the Gini mean difference in the following way. Assuming that the means of the groups are given, it is reasonable to postulate that the between-groups component should not change simply because of the degree of within group variation. They state, "It follows that the between groups component in the general case is the same as the between groups component in the special case where within group variation is zero for every group" (p. 150). Hence the between-groups concentration curve can be constructed from the cohort mean values (μ_i) and their population shares (p_i). The between-groups Gini coefficient is defined as $\Delta_B/2\mu$ where

$$\Delta_B = \sum_{i \neq j} p_i p_j \mid \mu_i - \mu_j \mid$$

They go on to partition the total Gini mean difference in terms of expected values:

$$(1) \quad \Delta = E \mid x_i - x_j \mid = \sum p_i^2 \Delta_i + \sum_{i \neq j} p_i p_j [E \mid x_i - x_j \mid]$$

where x_i and x_j are statistically independent observations and Δ_i is a cohort Gini mean difference. The within-groups component of the Gini mean difference is $\Delta - \Delta_B$:

$$(2) \quad \Delta_w = \sum p_i^2 \Delta_i + \sum_{i \neq j} p_i p_j [E \mid x_i - x_j \mid - \mid \mu_i - \mu_j \mid]$$

Bhattacharya and Mahalanobis conclude that while one cannot directly draw up a

*Professor of economics and urban studies, Portland State University.

concentration curve of overall within-group inequality, as one can for the between-group differences, the area between the latter curve and the L-curve "indicates the effect of within group disparities" (p. 151). This is the same procedure which I used to derive the P-Gini as a measure of within-cohort inequality effects. Yet Nelson cites this article in his opening argument to prove that I incorrectly disaggregated the Gini. We should note that the second component of equation (2) will be zero if within-cohort distributions do not overlap. Since considerable income variation occurs within cohorts, overlaps prevail and hence

$$(3) \sum_{i \neq j} p_i p_j \,|x_i - x_j| > \sum_{i \neq j} p_i p_j \,|\mu_i - \mu_j|$$

The magnitude of the interaction effects is measured by the difference between the terms in the above inequality, and is directly related to the degree of within-cohort inequality, given the means and population shares of the cohorts. But to pinpoint the economic meaning of the interaction effects contained in the within-group coefficient, it will be helpful to shift to the game-matrix framework recently developed by Graham Pyatt.[1]

Picking up the expected value approach mentioned above, Pyatt utilizes a variant of the Gini formula in which the average expected gain replaces the average of absolute differences; this reduces the numerator by one-half, and the denominator is accordingly changed from 2μ to μ, thus

$$(4) \quad G = (1/n^2) \sum_{i=1}^{n} \sum_{j=1}^{n} \max(0, x_i - x_j) \cdot 1/u$$

where $\max(0, x_i - x_j)$ stands for the higher of the two things within the parenthesis. The numerator of the Gini ratio now represents the average of expected gains of all individuals when placed in a game setting in which each draws the incomes of other persons at random, but keeps his own income if the income drawn is of lower value. Hence the negative $(x_i - x_j)$ values are recorded as zeros (and included in the average); obviously no one can lose in this game. The Gini coefficient can now be interpreted as the average expected gain (in an income comparison game) expressed as a percent of the mean income. Furthermore, the elements of the coefficient have a clear interpersonal reference. If the expected gain for an individual is high, he feels depressed because he would likely be better off in someone else's shoes; if it is low, he feels satisfaction at having moved ahead of most others in the economic race. In an unrestricted game setting involving the whole population the richest consuming unit would have an expected gain of zero and the poorest (with no income) would have an expected gain equal to the mean income. In a society with groups, in our instance age cohorts, we can meaningfully disaggregate the total of expected gains by using the matrices shown in Table 1. This approach also allows us to show conditional expectation or the average expected gain for individuals in group i drawing at random incomes from group j. This can be stated as

$$(5) \quad \text{average expected gain} = \sum_{i=1}^{k} \sum_{j=1}^{k} E(\text{gain} \mid i \rightarrow j) p_i p_j$$

with p values representing the population shares. A matrix E of all conditional between-groups and within-groups expected gains, normalized by dividing each row by the mean income of the respective population group, is presented in Table 1A for the 1974 family income distribution. In this normalized matrix E^* there is shown in each row the expected gains for an age cohort vis-à-vis other cohorts as well as the within-group gain; all are expressed as a percent of the mean income of the age group listed in the left column. The elements which constitute the main diagonal of the matrix are the cohort Gini coefficients. The values in the right-hand column marked E^*p show the average normalized expected gain for the members of each age group. To derive these average values each

[1] I will in general follow Pyatt's notation. My presentation in the next few pages is mainly a summary and application of his innovative analysis to the explication of the P-Gini.

TABLE 1—*U.S.* FAMILY INCOME DISTRIBUTION, 1974

A. Lorenz-Gini, Normalized Expected Gains E^*

Age Groups	14-24	25-34	35-44	45-54	55-64	65 +	Population Shares p	E^*_p
14-24	.3254	.6999	.9506	1.1046	.9068	.4096	.0758	.7922
25-34	.1161	.2890	.4266	.5138	.4137	.1726	.2281	.3506
35-44	.0784	.2013	.3054	.3716	.2984	.1230	.1948	.2504
45-54	.0641	.1656	.2543	.3103	.2491	.1026	.2016	.2083
55-64	.0996	.2416	.3543	.4248	.3433	.1451	.1554	.2916
65 +	.3174	.6690	.8983	1.0384	.8558	.3882	.1442	.7500

B. Expected Gains, *P*-Equality Conditions E^*_2

Age Groups	14-24	25-34	35-44	45-54	55-64	65 +	E_2p
14-24		.5230	.8087	.9779	.7340	.0699	.5982
25-34			.1876	.2987	.1386		.1183
35-44				.0936			.0189
45-54							
55-64			.0431	.1406			.0368
65 +		.4234	.6905	.8486	.6207		.4987

C. P-Gini, Expected Gains E^*_1

Age Groups	14-24	25-34	35-44	45-54	55-64	65 +	E^*_1p	Income Shares π
14-24	.3254	.1769	.1419	.1267	.1728	.3397	.1940	.0475
25-34	.1161	.2890	.2390	2151	.2751	.1726	.2323	.2174
35-44	.0784	.2013	.3054	.2780	.2984	.1230	.2315	.2205
45-54	.0641	.1656	.2543	.3103	.2491	.1026	.2083	.2496
55-64	.0996	2416	.3112	2842	.3433	.1451	.2548	.1686
65 +	.3174	.2456	.2081	.1898	.2351	.3882	.2513	.0965

Source: *Current Population Reports*
Note: Inequality coefficients derived from these matrices are equivalent to graphic straight line approximations and therefore are slightly understated

element in row i is weighted by the fractional population share p_j of the appropriate column. Note that the E^*p values are highest at each end of the age curve and lowest at the center; this in part reflects the fact that the groups with the lowest mean incomes have the most to gain and in part that the gains have been normalized by the separate group means rather than by a single population mean. Finally, we arrive at the overall Gini by taking the weighted average of the E^*p values, the weights being the income shares (π):

$$G = \pi E^*p = .343 \tag{6}$$

E^*_2 in Table 1B represents the normalized expected gains for the same distribution on the assumption that the members of each cohort have incomes centered at the mean value of the cohort; this is the *P*-equality condition. The diagonal elements are zero because we specified no inequality within cohorts. The off-diagonal (i, j)th element where mean group $i >$ mean group j will be zero because there is no possibility of a member of a richer group drawing a higher income from a poorer group. (Note however that in the E^* matrix this does occur because the distributions overlap.) As before, we can derive the Gini,

(7) Age-Gini $= \pi E_2^* p = .113$

This coefficient represents the average expected gain (relative to the mean income) which would exist under *P*-curve equality conditions. With the usual 45 degree line standard, all differences in income are regarded as alike; under the *P*-curve standard, differences related purely to the stage in the life cycle are considered normal, functional, and compatible with long-term equality conditions. Hence in E_2^* the youngest age group would still show an expected gain of .523 when comparing themselves with the 25-34 cohort, but the latter group shows zero gain vis-à-vis the younger cohort.

We can now derive the matrix of income differences with reference to the *P*-curve standard by subtracting E_2^* from E^*. This yields the *P*-Gini matrix E_1^* shown in Table 1C. The *P*-Gini coefficient is $\pi E_1^* p$ with a value of .230, but the crux of the matter lies in the interpretation of the elements of this coefficient and its meaning as an inequality index. The diagonal elements of E_1^* are the same as in E^* but the off-diagonal elements of E_1^* now represent the interaction effects, and these require examination. Let us look first at the (i, j) elements where mean $i >$ mean j. In the *L*-Gini E^* matrix, these elements represent expected gains when members of a lower income cohort (say the 25-34) compare themselves with a higher income age group (35-44); the expected gain shown is .4266 times their income. In the E_1^* matrix this is reduced to .2390 by the normal expected gain (.1876) associated with this age difference as shown in E_2^*. To put it in microeconomic terms we can think of 30-year olds comparing themselves not only with each other but with 40-year olds as well. If the 40-year olds have the "normal" incremental increase, there will be no sense of inequality generated within the younger group. But if a large degree of income variation exists in the 40-year cohort, this will also affect the younger group (and all other cohorts). A family in the 40-age group with an income of twice the mean for the group will generate feelings of inequality among members of other cohorts beside his own. This will be reflected in higher expected gains for the 30-year old cohort (vis-à-vis the 40-year olds) which are attributable to the inequality in the 40-year cohort, and further amplified if there is inequality in both cohorts. It is this increment of expected gain generated by the overlapping distributions which appears in the E_1^* matrix (for example, row 2, col. 3). However, a large part of the actual income difference between the two groups is tempered by the recognition that it is normal and transitory (for the younger group); hence the inequality contributed by such off-diagonal elements will typically be smaller than the diagonal elements. This corresponds to psychological reality. Sensitivity to income differences is more intense among members of the same age cohort than with those older and further along the life cycle curve. The differences within cohorts are not buffered by what are regarded as normal stage-of-life cycle income increments. The remaining effects of the overlap are shown in the corresponding (j, i)th element, row 3, column 2. Note that in E_2^* this cell is empty since there is no expected gain when mean $j >$ mean i. But inequality in the 40-year cohort results in families with incomes lower than some of those in the younger group, and this will be reinforced by inequality within the younger group pushing some families well above the lower incomes prevailing in the older cohort. These overlaps create an expected gain for some 40-year olds in the pair-wise income comparisons with the members of the younger cohort. The full value of this expected gain as it appears in E^* (.2013) is transferred to E_1^* since there would be a zero in this cell under *P*-equality conditions, and thus the gain is entirely attributable to inequality within groups, not to the difference between group means. The psychological reality underlying this interaction term is quite evident: it is illustrated by the "young upstart" phenomenon, felt most keenly by the lagging 40-year olds who resent seeing some 30-year olds moving ahead of them on the income ladder. But since this is experienced by only a limited proportion of the older cohort, the average expected gain for the whole cohort may be rather small.

To summarize, the E_1^* matrix contains all

the elements of expected gain that can logically be attributed to within-cohort inequality. The remaining elements of the Gini coefficient are those arising from mean differences between cohorts; these mean differences are regarded as normal, functional, and consistent with long-term equality of incomes; they are therefore included in the equality standard described by E_1^*.

Now let us return to Nelson's criticism. When he states that a proper or true estimate of the *P*-Gini is given in his equation (4), he is in effect looking at our E_1^* matrix, selecting just the diagonal elements $(\Sigma\lambda_i\gamma_iG_i)$ and tossing the off-diagonal interaction terms into the waste basket as having no significance for an inequality measure based on the *P*-equality standard. Since there are k^2 elements in this matrix, each weighted by population and income shares, the decision to throw out all but the k diagonal elements yields a suspiciously low percent of the total Gini attributable to within-age group inequality effects. To compensate for the excluded terms, Nelson blows up the weights for the diagonal terms by using either population shares *or* income shares. His exclusion of the interaction terms in E_1^* is equivalent to putting blinders around members of each cohort so that they can only compare themselves with others in the same cohort; hence they will be insulated from the effects of inequality in other cohorts, as well as the muting effects of the normal difference components, which as we have shown do influence their perception of inequality. Cutting groups off from each other violates the essential meaning of the Gini formula which specifies unrestricted pair-wise comparison of incomes.

Finally, if we accept the *P*-curve standard of equality, Nelson's alternate *P*-measure yields an incorrect index of deviation from this standard. Referring to 1972 data from his Table 2, we find that the single weighted cohort Gini is .333 compared with an *L*-Gini of .359. From this it would be natural to assume that if we eliminated all intracohort inequality the Lorenz curve would move inward by .333 to a residual Gini of .026. Instead we find that elimination of intracohort inequality moves the *L*-curve inward by only .239 points, leaving a residual Gini of .12; this is the value given by the age-Gini.

Why are there such differences in our estimates of the trend of inequality? I look at the value of $\pi E_1^* p$ in 1947 and in 1972 and note a decline of .064 or 23 percent in the *P*-Gini ratio. Nelson looks at just the diagonal in E_1^* and finds that if one uses the weights specified by the Gini formula $(\lambda_i\gamma_iG_i)$ this segment of the matrix has declined by 14.8 percent (his Table 3, line 4, "total"). If he "standardizes" with one set of weights taken from 1947 and applied to both years then the decline is only 6.6 percent.[2] I have given reasons for using the whole of matrix E_1^* in a world where the life cycle income curve has become arched, and where recognition of the functional nature of life cycle differences has been embodied in the equality standard. I suggest that to use Nelson's approach one must posit an alternative paradigm, namely a world in which life cycle income curves have become straight lines with a zero slope (as shown in Figure 1A) and the distribution characterized only by within-cohort inequality. The E^* matrix of this distribution would be the same as E_1^*. Since there are no between-cohort differences, the diagonal elements of the matrix would become a fair proxy for all the elements of the matrix. When we extend the cohort Gini approach to societies in which the age-income curves are highly arched (as shown in Figure 1B) we imply that the curvature makes no difference and we can abstract from it by taking many segmental snapshots (age-cohort Gini ratios) and ignore the extent of differences between the means of these groups, the interaction terms, and other related aspects of the distribution. We can see the im-

[2]The 14.8 percent (or the 6.6) average decline in inequality is low in part because the decline in five cohorts was partially offset by a 21 percent increase in inequality for the 14–24 age cohort. The increase in inequality occurred during a period of great expansion in subsidized opportunities for higher education. This shows that an egalitarian measure to broaden lifetime opportunities may produce a greater variance in initial income positions. Static cohort measures don't reveal this.

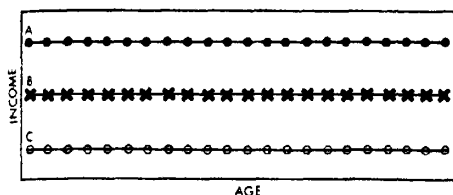



FIGURE 1A

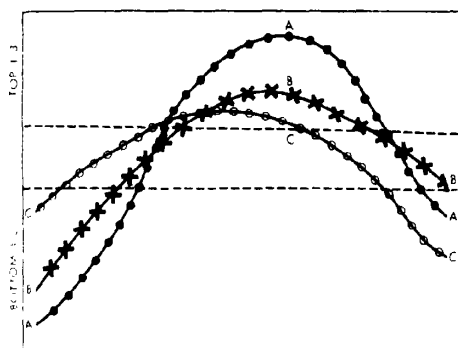



FIGURE 1B

portance of these other aspects by comparing Figures 1A and 1B.

The world of Figure 1A is one with less Lorenzian inequality than 1B but with a much higher probability of permanent stratification; those in the bottom third of the income distribution tend to stay there over their lifetimes, and the same applies to the middle and the top of the distribution. Now compare this with Figure 1B which shows income curves that roughly typify our income distribution. Note that a marked curvature in the age-income profile produces mobility with reference to income levels and population rankings. This is reflected in the statistics on family incomes. In 1974, families headed by a high school drop-out (age 45–54) averaged $15,042; this placed them for a time in the second from the top quintile. Furthermore, since 44 percent of this group had incomes above the group mean, a large number of such families experienced the status that goes with being in the top quintile for part of their lives. This age-related mobility, induced by highly arched age-income profiles, is accompanied by a second type of mobility, namely, movement in the rank ordering within one's cohort as the cohort moves across the age-income profile. Insofar as there are many curve types which cross each other, there will be cohort mobility. This is shown in Figure 1B by the change in letter rankings at different ages. These differences in the two types of distributions are not reflected in the weighted age-cohort Gini. If this coefficient has the same value in Figures 1A and 1B, the perceived level of inequality and lifetime inequality will be lower in the society with the arched profile, even though its *L*-Gini is higher. Hence it is appropriate that the *P*-Gini indicates a lesser degree of inequality than the weighted cohort Gini as we move from a flat to a curved age-income profile. The *P*-Gini captures the age-related mobility effects in the values of the off-diagonal elements of the E_1^* matrix where income differences between cohorts are modified by the expectations of life cycle mobility. Certainly the *L*-Gini which lumps all income differences in the same box is an inadequate measure of both current and historical inequality.

The measures can be differentiated in terms of the Pyatt game framework. The *L*-Gini may be interpreted as the average expected gain for the population under conditions of an unrestricted pair-wise income comparison game, with the gain being expressed as a percentage of the mean income. The *P*-Gini uses the results of this game, but rather than comparing the score with equality defined as zero gain it is compared to a score which emerges when average stage-of-life cycle income differences are postulated as an equality condition. The weighted cohort Gini attempts to measure the expected gain that would exist in a world where stage-of-life cycle income differences have been eliminated but where inequality within cohorts remains; it crudely approximates this condition in the general case by calculating the average expected gain for a series of restricted pair-wise comparison games in which members of a cohort are paired only with others in the same cohort. Since I believe that average income differences between cohorts are normal and functional, and their elimination unneces-

sary as a long-run equality condition, I find the specifications of the *P*-Gini game more interesting and meaningful.

William Johnson's paper raises the issue of inequality of permanent or lifetime incomes. While I disagree with his main conclusion, I fully agree that the *P*-Gini does not indicate actual differences in lifetime incomes. However, I did make some qualifications to this effect in my paper: the *P*-Gini "... more closely approximates a measure of long-run interfamily inequality" than the *L*-Gini (p. 601). This was followed by a discussion of the difficulties of measuring the inequality of actual lifetime incomes, particularly under conditions of economic growth (pp. 601-02). I should have qualified this further; no equality measure based on cross-section data for a particular year can be used to measure actual lifetime income differences. What my measure shows is that only a part (two-thirds) of the total area of inequality need be eliminated in order to arrive at conditions which will produce equality of lifetime incomes (under stationary assumptions); and currently about one-third of the Lorenzian inequality of annual incomes would remain under long-term equality conditions because it represents age-related income differences quite compatible with this goal. We should note that the frame of reference here is the partitioning of a single year's inequality measure into components which are consistent or inconsistent with long-term equality, not the actual measurement of long-term incomes. We can make some inferences from this, but they should be carefully qualified.

Johnson's statement "... that Paglin's adjusted Gini coefficient will always underestimate the true extent of lifetime income inequality" will be shown to be false. I go along with two of the assumptions of his model, namely, a stationary state and stable population cohorts. However, Johnson's results also depend on the crucial though unmentioned assumption that each person moving along his lifetime income track always keeps the same rank order within his cohort. This is equivalent to assuming no crossing of life cycle income curves or zero mobility within age cohorts. Otherwise his equation (7) will overestimate lifetime income inequality. This can be easily seen in Figure 1B. Lifetime income differences for *A*, *B*, and *C* will be far smaller than indicated by taking the absolute differences between the three curves at each age level as Johnson specifies in equation (7). Since the *L*-Gini, age-Gini, and *P*-Gini are all calculated at a point in time, and are unaffected by who occupies a particular rank in the income distribution, they will show no reduction in inequality as a result of intracohort mobility. Therefore, the *P*-Gini will yield an estimate which is above or below the lifetime income inequality depending on the degree of cohort mobility. Since Johnson's proof depends on zero mobility within cohorts, it will be of interest to review the data in this area.

Bradley Schiller's recent research documents the surprisingly high degree of mobility which exists within cohorts. Schiller's study is based on the incomes of over 74,000 male earners tracked for a period of 14 years. Using longitudinal earnings data drawn from workers covered by social security legislation, Schiller ranked workers who were age 30-34 in 1957 by income levels, and ranked them again on the basis of their 1971 earnings. The workers were classified on a 20 interval ranking scale, with each "ventile" containing 5 percent of the sample. Hence the first ventile included the top 5 percent of the income recipients, etc. Even though the study started with an age cohort which is marked by more stability than the younger cohorts, Schiller found a remarkable degree of mobility within this cohort. Defining mobility as a move of two or more ventiles, 71 percent of the cohort was mobile, the average move "... was 4.22 ventiles (21 percentiles) up or down the earnings distribution, or over one fifth of the way from one end of the distribution to the other. Hence mobility of relative status not only is a common experience, but also involves very large movements" (p. 115). Furthermore, mobility was evident all across the income distribution. Over two-thirds of those in the bottom ventiles moved upward by more than four ventiles while 60 percent of those in the top three ventiles moved downward over four

ventiles. (Since the cohort was in a sharp upswing phase of the life cycle income curve, most of those who moved downward in ranking nevertheless moved upward in real income; the slopes of their curves were just not as steep as those who previously ranked below them.) Schiller's apt conclusion is that even if the income chairs in the distribution are relatively fixed, the persons who occupy those chairs are highly mobile.

I believe these results have important implications for the trend of inequality. Mobility reduces the dispersion of lifetime incomes much below the annual income estimate. Mobility is a function of the crossing of individual life cycle income curves. With a relatively flat average age-income profile, individual lifetime income paths are less likely to cross than in a society with a highly arched average profile where there is a greater range in the individual slope coefficients and curve types. As I have shown (p. 600), our society has experienced a noticeable increase in the degree of curvature in the average age-income profile. If this is accompanied by increasing mobility within cohorts, then measures such as the weighted cohort Gini will not only overestimate inequality in a particular year (as pointed out above) but will underestimate the decline in inequality. While the *P*-Gini adjusts for average age-related inequality it also fails to catch the accompanying intracohort mobility. Until we are able to modify our static inequality coefficients by an index of mobility, or collect more longitudinal household income data for an extended period of time, our estimate of inequality of lifetime incomes (or the more difficult trend in the inequality of lifetime incomes) will remain crude. However we do know the direction of the bias in the conventional coefficients.[3]

The paper by Sheldon Danziger, Robert

[3]This is beginning to show up in the University of Michigan longitudinal income data. The Gini ratio for family incomes in 1973 was 9.9 percent higher than the Gini of the seven year income average of these families. Using the Theil index, it was 24 percent higher, and for the variance of the *log* of income it was 26 percent higher. Note that seven years is only a small segment of the family life cycle (Saul Hoffman and Nripesh Podder, p. 339).

TABLE 2

	D-H-S Results		Corrected Results	
	1972	1965	1972	1965
L-Gini	.4043	.3885	.4140	.3885
Age-Gini	.2344	.2073	.1575	.1385
P-Gini	.1699	.1812	.2565	.2500

Source: D-H-S results are from their Table 1. The corrected age-Ginis are derived from the D-H-S data but use the proper formula. My 1972 *L*-Gini is based on *CPS* computer tapes and is taken from Taussig.

Haveman, and Eugene Smolensky (hereafter referred to as D-H-S) is a wide ranging one, but its main theme is the supposedly eccentric counterintuitive performance of the *P*-Gini as illustrated by the statistics in their Table 1. I can be fairly brief in my reply to the long discussion of the misbehaved *P*-Gini because their calculation of the age-Gini and *P*-Gini coefficients are simply erroneous. These large errors are shown in my Table 2. The reasons for the differences are not data or computer errors, because D-H-S generously provided me with their data decks. This was of particular importance since for 1965 they used the unpublished Survey of Economic Opportunity (*SEO*). Surprising as it sounds, they simply calculated a different measure! The source note under their Table 1 and the details spelled out in footnote 2 reveal the misspecification of the age-Gini and hence the *P*-Gini: "Paglin's Ginis are based on a division of families into 6 age cohorts. We classify all household units into 24 mutually exclusive cohorts by type of living unit (family or unrelated individual) and sex of head, in addition to the 6 age classes used by Paglin." My age-Gini is based on the mean income differences between age groups. Their age-Gini is really an age-demographic Gini with four different means in each age group and thus represents some intracohort variation as well as difference between cohorts. The comparable *P*-curve or equality standard implied by their revision is that existing mean differences between male-headed and female-headed households are consistent with long-term equality. (This is an odd anti-egalitarian standard to offer as an improvement.) I

think it should be clear that stage-of-life cycle differences wash out in the long run because the aging process has no barriers other than premature mortality, and hence it is appropriate to include such differences in the equality standard. The same cannot be said for income differences based on sex, or the age-schooling standard used in the paper by Joseph Minarik.

The differences between the D-H-S ratios and the corrected versions can be seen in Table 2. Their age-Ginis are about 49 percent larger than those based on a proper computation, while their *P*-Gini coefficients are 38 and 44 percent too low (for 1965 and 1972, respectively). Perhaps more significant is the reversal of the trend of the *P*-Gini on which D-H-S placed such importance. It now moves in the same direction as the Lorenz-Gini and thus according to their criteria of a well-behaved inequality coefficient should receive approval. However, I believe these results (no matter how they come out) prove very little. The 1965 data were drawn from an *SEO* sample while the 1972 data are from the Census *Current Population Survey (CPS)*.[4] Furthermore, their manipulations of the data fail to consider the totality of the E_t^* matrix on which the *P*-Gini is based.

Leaving aside the sampling issue, it would be fairly easy to find two years (well-spaced apart) in the *CPS* family income series in which the *L*-Gini has moved up and the *P*-Gini down. This results from the greater downward trend in the latter series. If we take first differences, the two coefficients have a remarkably high degree of covariance. This is shown in Table 3 where in 20 years out of 23 both coefficients changed in the same direction, while in 1 year there was a change in one coefficient and no change in the other. In terms of short-run changes in inequality due to such factors as unemployment, what is "Lorenz-efficient" also turns out to be "Paglin-efficient." D-H-S claim that an increase in

TABLE 3—GINI COEFFICIENTS, CHANGE FROM PREVIOUS YEAR (Δ_t)

Year	*L*-Gini	*P*-Gini
1948	-.009	-.010
1949	.010	.010
1950	-.004	-.005
1951	-.014	-.009
1952	.013	.007
1955	-.007	-.011
1956	-.011	-.006
1957	-.004	-.010
1958	.003	-.001
1959	.012	.015
1960	.003	.002
1961	.007	.008
1962	-.011	-.021
1963	-.005	-.003
1964	-.008	-.006
1968	-.012	-.012
1969	.005	.000
1970	.007	.008
1971	.001	-.004
1972	.003	.002
1973	-.003	-.003
1974	-.001	-.003
1975	.003	.005

Source: The author, Table 3, updated to 1975.

teenage unemployment will increase the age-Gini and hence possibly lower the *P*-Gini. I am not sure this will happen. Increased unemployment also increases inequality within the cohort; the *L*-Gini responds both to changes within and changes between cohorts. Therefore, it is quite possible that the *L*-Gini will increase more than the age-Gini, thus showing an increase in in the *P*-Gini. However reality is usually more complex than synthetic examples. Teenage unemployment affects not only the young household cohort, but since many teenagers are second or third earners in families headed by persons in middle age, the effects are diffused across the whole distribution, and therefore cannot be isolated. However, for unemployment in general the *P*-Gini performs quite well. With trend removed, I found that the *P*-Gini typically changes in the same direction as the unemployment rate; this is also confirmed by Minarik (fn. 5).

D-H-S demonstrate in Section IV and Table 2 that certain transfers are more Lorenz-efficient than Paglin-efficient. The

[4]The *SEO* data for 1965 show a lower *L*-Gini than the comparable *CPS* data for families and single individuals combined: .3885 versus .403. The latter figure is from Michael Taussig and is calculated from computer tapes.

implication seems to be that the *L*-Gini is superior to the *P*-Gini because the former will sometimes show a larger change for a given size of income transfer. I don't believe social policies (or inequality coefficients) should be evaluated in this way. Transfers between age cohorts have different effects on long-term inequality than reductions of inequality within cohorts. The *L*-Gini lumps these together; this is not desirable.

D-H-S question my use of Census *CPS* data to construct the *P*-curve: "Because public transfers are included in this income definition, Paglin has rejected the optimality of the age-income profile generated by the market. Yet no explanation is given ..." (p. 511). The market system generates private incomes and private goods. We have seen fit to modify this result through income transfers which may be regarded as public goods; these are provided through a political decision-making process. Insofar as people express their preferences through both mechanisms there is nothing wrong in basing an equality standard on the resulting distribution. I believe the *CPS* income concept should be broadened to include those in-kind transfers which are common elements in all consumer budgets, but I am constrained to work with what is available. I also do not accept the D-H-S strictures on my use of income data which includes social security payments. These represent political decisions to alter the time shape of our incomes. Why shouldn't the *P*-curve norm reflect these decisions? I never stated that the age-income profile should be immune from all political decision processes. That is their interpretation. However, there is very little theoretical support for an equality standard which requires a flat profile as specified in the Lorenz curve. D-H-S seem to concur in this judgment. Their analysis (Section III) of the determinants of the age-income profile exhibited in their Figure 1 clearly points to one conclusion: If such basic factors as "the returns to investments in human capital," "the distribution of inherent physical and mental capabilities by age," and the "earnings effects of labor-leisure choices by age" lead to a curved age-income profile, then there are excellent reasons for not violating such powerful forces by setting up a dysfunctional flat equality standard that is also unnecessary to achieve long-term equality. D-H-S concede that "An inequality measure which allows for life cycle variations is appealing" (p. 512), and they acknowledge "... the weaknesses of the Lorenz-Gini as a standard of equality since it does not allow for life cycle income variation" (fn. 5). What then are their objections? Mainly, I think, that the *P*-reference curve and hence the age-Gini will rely "on annual observations of an arbitrarily observed pattern" (p. 512), and the presumed perverse short-term changes in the *P*-Gini. The latter judgment, as I have noted, was based on a statistical misspecification of the age-Gini. Certainly their listed determinants (*A* through *G*) of the *P*-curve standard are all likely to act slowly and gradually. This is exemplified by the high degree of correlation in the short-term swings of the two inequality coefficients (my Table 3) indicating that the main difference between the two measures is in their level and trend. To get around what they consider an arbitrarily based standard, they would replace the *P*-curve standard derived from average age cohort incomes with one which depended exclusively on "... a collective judgment on age-related need ... analogous to the official *U.S.* poverty definition" (fn. 5). I am not optimistic about the validity of such judgments, especially as we are at a level well above the poverty requirements. I would say the collective judgment is better revealed by the average age-income statistics which recognize all market expressed preferences, productivity considerations, and political decisions, than by the judgment of a team of experts. The life cycle issue is not limited to needs only, and that is also true of the equity dimension. Finally, D-H-S refer to George Garvy who in 1952 raised the issue of an age-based standard but rejected it because "the subtraction of Gini coefficients from separate underlying distributions is uninterpretable" (fn. 13). As I have shown, a meaningful disaggregation of the *L*-Gini based on the *P*-curve standard can be made, and the net *P*-Gini has a clear interpretation in expected gain terms.

Joseph Minarik's paper attempts to extend the equality standard to include educa-

tion as well as age. Its major fault as I indicated earlier is that the aging process is perfectly democratic, but higher education may still have elements which John Kurien would refer to as opportunity-related rather than choice-related. The age-schooling standard may thus be questioned as a legitimate specification of equality since the age-schooling Gini contains income differences which wash out in the long run (age) and education-related differences which don't. Hence the residual Gini is not very useful for analyzing the trend of inequality which Minarik oddly enough finds the only interesting application of an inequality measure. The disaggregation of the *L*-Gini in a current year into an age-Gini component and a residual is meaningful because it points to the dual set of forces at work in producing the total matrix of inequality interactions; policy makers should be cognizant of these differences. The same method of partitioning inequality can also be extended to show the conflicting nature of various equality standards, for example, equality of earnings versus equality of family incomes. If we postulate perfect equality of earnings but allow the existing variation in the number of earners per family to remain, the 1974 *L*-Gini of family incomes would still be .265 or 74 percent of the actual inequality level. (I assumed that families with no earners would receive transfer incomes equal to three-fourths of the income of the single earner families.)

I do not share Minarik's surprise over the barely noticeable rise in the *P*-Gini from 1967 to 1974 (a difference of .005) when earned income only is used (his Table 2, col. 4). Economic theory points to this as an expected result of the expansion of our transfer programs. Transfer income has in part been substituted for earned income in many households which face high marginal tax rates. Minarik pulls out transfers from the income distribution and assumes that the residual tells us what the trend in the market distribution of incomes would look like in the absence of transfers. Cash transfers and in-kind benefits are triggered by earned income levels; it is perfectly rational for some low income families to avoid entering the labor market when the loss in transfer benefits might offset the potential earnings. Given the nature of our transfer programs, and their rapid growth, Minarik's results are to be expected, but why draw gloomy conclusions on the trend of inequality from such data? An egalitarian who supports more generous transfers to reduce total income inequality must face up to the impacts which these transfers will have on the earned income distribution.

John Kurien suggests that an equality standard should allow for choice-related differences in incomes and provide a net measure of opportunity-related differences. This seems reasonable, but it fails as an operating principle because of the complex interconnection between choices and opportunities. Consider the question of family size and how we might adjust for it. Family *X* decides to have four children; family *Y* decides to have one child and spend more on education and travel. Assume both families have the same incomes. Since this is a choice-related state of affairs, Kurien would see no need to correct for size of family in evaluating the real welfare levels of the two families. From this view, children are regarded as consumer goods comparable to luxuries and travel. In a way they are, but if we now consider the children as independent elements in the welfare equation, our point of view must change. The children in the large family may have to spend more time working as teenagers and less may be provided for them by way of higher education. The choices of the parents have become the opportunity-related conditions of the children. In policy terms, do we provide subsidies to the children of large families to offset their more limited opportunities which result from the choices made by their parents? From the other point of view, should we require small families to subsidize the choice-related decisions of the large families? Kurien's partitioning principle doesn't help us much.

There are a number of other income-related factors associated with the life cycle, particularly the productivity and work-leisure variables, for which there is no agreement as to how adjustments should be made. For all these, the use of the average life cycle income curve becomes a useful

though crude substitute for more explicit treatment. When millions of household units, differing in myriad ways, are all thrown into the same box for comparison in terms of a single variable (income), only limited adjustments can be made for the diversity of their needs, productive contributions, and circumstances. Hence, all equality standards tend to be simplistic, and the *P*-curve standard is no exception.

I won't try to recapitulate the themes of this extensive discussion; rather, I will emphasize just two points. The weighted cohort Gini, which we have traditionally employed to correct for the upward bias in the *L*-Gini, abstracts from the significant changes in life cycle income patterns which have occurred in the last several decades. Hence it fails to catch the effects of life cycle mobility and cohort mobility on inequality. These effects are important both for our current perception of inequality and for our estimate of lifetime income differences. Perceived inequality is not just a function of where we now are in the income ranking, but where we expect to be. An inequality measure should not abstract from the life cycle income curve but include it in the assessment process. In a limited sense, the *P*-Gini uses the expectation of average life cycle mobility in its income comparison matrix, and this may be regarded as a first step toward the inclusion of expectations in a traditionally static measure of inequality.

REFERENCES

N. Bhattacharya and B. Mahalanobis, "Regional Disparities in Household Consumption in India," *J. Amer. Statist. Assn.*, Mar. 1967, *62*, 143-61.

S. Danziger, R. Haveman, and E. Smolensky, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 505-12.

S. Hoffman and N. Podder, "Income Inequality," in Greg Duncan and James Morgan, eds., *Five Thousand American Families—Patterns of Economic Progress*, Ann Arbor 1976, 333-56.

W. Johnson, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 502-04.

C. J. Kurien, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 517-19.

J. J. Minarik, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 513-16.

E. Nelson, "The Measurement and Trend of Inequality: Comment," *Amer. Econ. Rev.*, June 1977, *67*, 497-501.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598-609.

G. Pyatt, "On the Interpretation and Disaggregation of Gini Coefficients," *Econ. J.*, June 1976, *86*, 243-55.

B. Schiller, "Equality, Opportunity, and the 'Good Job'," *Publ. Interest*, Spring 1976, *43*, 111-20.

M. Taussig, "Trends in Inequality of Well-Offness in the United States Since World War II," spec. rept., Inst. Res. Poverty, Univ. Wisconsin 1977, forthcoming.

U.S. Bureau of the Census, *Current Population Reports*, Series P-60, No. 101, Washington 1976.

NOTES

Nominations for AEA Officers

The Electoral College on March 18 chose Robert Solow as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1977. Other nominees (chosen by the nominating committee) are: for Vice-President (two to be elected), Edward Denison, Joseph Pechman, William Vickrey, Phyllis Wallace; for member of the Executive Committee (two to be elected), Robert Clower, William Nordhaus, Lester Thurow, Marina Whitman

Under a change in the bylaws as described in the *Papers and Proceedings* of this *Review*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices For the purpose of circulating petitions, address labels will be made available by the Secretary at cost

1978 Nominating Committee of the AEA

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Jacob Marschak has appointed a Nominating Committee for 1978 consisting of Andrew F Brimmer, Chairman, Carolyn Shaw Bell, Peter A. Diamond, John G Gurley, Robert M Haveman, Bert G Hickman, and Vernon L. Smith Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-elect, the Committee shall include any other member of the Association nominated by petition including signature and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1 No member of the Association may validly petition for more than one nominee for the Committee The names of the committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

On March 18, 1977 the Executive Committee of the American Economic Association voted to increase membership dues and subscriptions 5 percent effective January 1, 1978, a complete list of the new rates follows. This action was taken under the provision of the bylaws permitting the Executive Committee to increase the dues schedule in proportion to the increase occurring in relevant wage and price indexes Association membership dues for 1978 are as follows:

$26.25 for regular members with rank of assistant professor or lower, or with annual income of $12,600 or less; $31.50 for regular members with rank of associate professor, or with annual income of $12,600 to $21,000; 36.75 for regular members with rank of full professor, or with annual income above $21,000, $13.00 for junior members (registered students). Certification must be submitted yearly. Subscriptions (libraries, institutions, or firms) are $39 35 a year. In countries other than the United States, add $3.70 to cover extra postage. Family member (*second membership without publications*, two or more persons living at same address) . . . $5.25.

The Asia Foundation has awarded the American Economic Association a grant to assist graduate students and visiting professors from Asia currently at institutions east of the Mississippi to attend the annual meeting of the Association The 1977 meeting will be held in New York City, December 28 30. Approximately seven travel grants up to $150 each will be awarded Write the American Economic Association 1313 21st Avenue South, Nashville, TN 37212 for an application form. To be considered for a travel grant completed forms must be returned by October 31, 1977

Scientists and Engineers in Economic Development (SEED) Program

The National Science Foundation, through a program funded by the Agency for International Development (AID), will provide support for individual *U S* scientists and engineers to apply their experience to specific problems of developing countries The program's objectives are to (1) enable *U S* scientists and engineers to share experiences with their counterparts in developing countries through the conduct of specific research and education projects contributing to the economic development of the host country, (2) establish long-term collaborative relationships between *U S* and foreign institutions, and (3) increase the capability of scientific and technical institutions in developing countries to contribute to economic development Applicants are limited to scientists and engineers from *U S* academic institutions with at least five years of postdoctoral or equivalent experience in teaching or research and who will return to their institutions on completion of the project For guidelines for the preparation of proposals, write or call Division of International Programs, National Science Foundation Washington, DC 20550, Telephone (202) 634-7930

The Financial Accounting Standards Board invites research papers on the economic consequences of financial accounting standards and changes therein. Specific examples of potential economic consequences that are of interest to the Board can be obtained from George J Stabus, Director of Research and Technical Activities Financial Accounting Standards Board, High Ridge

Park, Stamford, Connecticut, 06905. The staff of the Board will select several of the papers for presentation at the Conference on the Economic Consequences of Accounting Standards to be held at the FASB office in March 1978. Deadline for papers is December 15, 1977.

To introduce members of the American Economic Association to *Economic Development and Cultural Change*, the journal is offering a special discount. Members who are not presently subscribers may enter, before July 1, 1977, one-year subscriptions for $12.00 (regular rate, $15.00). And please note, all subscribers are entitled to order an important new publication, *Essays on Economic Development and Cultural Change in Honor of Bert F Hoselitz*, at a 25 percent discount. For further information write to Orlie Higgins, The University of Chicago Press, 5801 S. Ellis Ave , Chicago, IL 60637.

Members of the American Economic Association insured under the Life Insurance Plan will receive a credit on their April 1, 1977 Notice of Payment Due equal to 50 percent of the amount they contributed during the policy year ending September 30, 1976 This credit was made possible by favorable experience and the continued growth of the Plan. These credits, of course, cannot be guaranteed Due to a change in the Wisconsin insurance regulations, effective April 1, 1977, Wisconsin members will now be eligible for benefits under the group plan (up to $72,000 for a member, $35,000 for a spouse and $2,500 for each eligible child subject to normal underwriting) Prior to this change, members living in Wisconsin have been issued individual life insurance policies with lower amounts of coverage available

All inquiries about the Life Insurance Plan or any of the other Plans in the Group Insurance Program should be made to the Administrator, 1707 L Street, N W , Washington, D C. 20036 Telephone (202) 296-8030

Southern Economic Association meeting, Nov. 2-4, 1977, New Orleans, LA.

The *Pomerance Prize for Excellence in the Area of Options Research* was established in December 1976. An award of $1,500 in cash will be given annually by the Chicago Board Options Exchange for the most outstanding completed study on the market for exchange-traded options Individuals who are employees or members of CBOE or are nominees of member organizations are not eligible. Studies which have been published prior to November 1976 are not eligible. Applications and submissions for the first award must be received by September 30, 1977. The recipient of the award will be announced December 30, 1977. To receive application forms and further information please contact: Michel Tremblay, Liaison—Pomerance Prize Committee, Chicago Board Options Exchange, 141 West Jackson Blvd., Chicago, IL 60604.

The Secretary of Labor's Invitational Conference on the National Longitudinal Surveys of Mature Women is scheduled to be held in Washington, D.C. January 1978. Scholars are invited to submit papers which deal with the employment experiences of mature women. The deadline is September 30, 1977. Information concerning criteria for selection and other communications should be addressed to Isabel V. Sawhill, The Urban Institute, 2100 M Street, N.W., Washington, D C 20037.

The Association for Comparative Economic Studies wishes to announce its new journal, the *Journal of Comparative Economics*, to be published quarterly by Academic Press, Inc., 111 Fifth Avenue, New York, New York 10003, beginning March 1977 Subscription rates may be obtained from Academic Press; members of ACES may subscribe through the Association's Executive Secretary, Elizabeth Clayton, Department of Economics, University of Missouri, 8001 Natural Bridge Road, St Louis, Missouri 63121. The editor invites submission of manuscripts for publication. Manuscripts (typed in triplicate) should be sent to John Michael Montias, *Journal of Comparative Economics*, Yale University, Box 16A, Yale Station, New Haven, CT 06520

The National Research Council will soon be conducting its biennial survey of a sample of Ph.D.s in the *U S* labor force The survey, sponsored by the National Science Foundation with additional support from the National Endowment for the Humanities and the National Institutes of Health has enjoyed a high level of response in 1973 and 1975. In early 1977 the third biennial survey will be a conducted to gather current employment and career information on a stratified sample of over 80,000 individuals who have received a doctorate in the sciences, engineering or the humanities between 1934 and 1976 The purpose of the survey is to obtain data which will help assess the status of the nation's Ph D.s and to develop policies and programs which affect this important segment of the population.

The Joint Committee on Research, sponsored by the Council on Foundations and the Foundation Center, is interested in contacting social scientists engaged in or planning research on philanthropy, including analyses of the history and current structure and functioning of philanthropic institutions and activities. The Committee's interest is to increase scholarly research on philanthropy by fostering communication among scholars working in this area, by reviewing proposals,

and by facilitating funding of those proposals that are of particular merit. The Committee is also interested in learning about instructional programs, including courses and seminars in graduate and professional schools on the topic of philanthropy. The Committee itself is not a funding agency, but it can be a useful information and referral source for scholars Members of the Committee are: William J. Baumol, Thomas R. Buckman, Fred C. Cole, James Stacy Coles, David F. Freeman, Marion Fremont-Smith, Robert F. Goheen, Barry D. Karl, Robert M. Lumiansky, Robert K. Merton, Eleanor B. Sheldon, John G. Simon, and Logan Wilson. Contact the Joint Committee on Research, Hugh F. Cline, Chairman, Joint Committee on Research, Educational Testing Service, Princeton, New Jersey 08540.

The jointly sponsored Ford and Rockefeller Foundation's Research Program is interested in receiving proposals focussing on the formulation, implementation and evaluation of population policy as it relates to social and economic development. Of particular interest to this year's program are proposals that may help in closing the gap between research and policy planning on development issues. Submissions are encouraged on a broad range of topics, including: determinants of demographic behavior, with emphasis on variables subject to planned intervention; consequences of population trends at the household, community or national levels; political, cultural, economic, and demographic factors influencing the formulation of population policies, demographic impact of public policies and programs, or conversely, the social, economic, and political impact of population policies.

The deadline for submission of proposals is July 1, 1977, and awards will be announced in December. The proposed research may begin on or after January 1, 1978. For further program information outlining application procedures, please write to: The Ford and Rockefeller Foundations' Research Program on Population and Development Policy, The Ford Foundation, 320 East 43rd Street, New York, 10017.

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Grants are likely to cover only lowest cost excursion fares and will rarely exceed 50 percent of full economy-class fares. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are ineligible. The deadlines for applications to be received in the office of the American Economic Association are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1, for meetings scheduled between March and June, November 1. Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, TN 37212.

Deaths

Taulman A. Miller, professor of economics, Indiana University, Jan 24, 1977.

Chester A. Phillips, dean emeritus, University of Iowa College of Business Administration, Dec 1, 1976.

Visiting Foreign Scholars

Patrick Messerlin, Université Paris XII. visiting assistant professor, University of Houston, Jan. 1977

Promotions

R. Keith Aufhauser· associate professor of economics, Queens College, Jan. 1, 1977.

William A. Bachman: assistant professor of economics, Niagara University, Sept. 1, 1976

Michael R. Dohan: associate professor of economics, Queens College, Jan. 1, 1977.

Kurt Hausafus· associate professor, Oberlin College

Michael J. P. Magill associate professor of economics, Indiana University, July 1976.

James H. Schulz professor of Welfare economics Heller Graduate School, Brandeis University, Sept 1976

Peter D. Sternlight senior vice president, Open Market Operations and Treasury Issues Function. Federal Reserve Bank of New York, Jan. 1, 1977

James M. Suarez. associate professor, Hunter College, Jan 1, 1977.

Administrative Appointments

Herman A. Berliner. assistant provost, Hofstra University, Dec. 1, 1976.

Bert T. Glaze director of evening school, Marietta College, Sept. 1976.

William Hamovitch: acting provost, Queens College, Feb. 1, 1977.

Rodney J. Morrison: chairman, department of economics, Wellesley College, July 1, 1977–June 30, 1980

R. Lynn Rittenoure, Naval Postgraduate School chairman, department of economics, University of Tulsa, July 1, 1977.

James E. Zinser: chairman, department of economics, Oberlin College, Feb. 1977.

Appointments

Sham L. Bhatia: assistant professor of economics Indiana University Northwest, Fall 1976.

Erwin A. Blackstone: associate professor, department of economics, School of Business Administration, Temple University.

Sandra Cohan, Federal Deposit Insurance Corporation: visiting associate professor, Oberlin College, Sept. 1976.

Dale S. Drum, Wichita State University: economist, Federal Reserve Bank of Chicago, July 15, 1976.

James M. Griffin, University of Pennsylvania: professor of economics, University of Houston, Jan. 1977.

Simon Hakim: assistant professor, department of economics, School of Business Administration, Temple University.

David T. Kresge: senior research staff, National Bureau of Economic Research, Jan. 1977.

Andrew Madsen: assistant professor, department of economics and commerce, Niagara University, Jan. 1, 1977.

Steven K. Palmer assistant professor of economics and management, Marietta College, Aug. 1976.

Claudio A. Pardo: economist, foreign research division, Federal Reserve Bank of New York, Nov. 4, 1976

Joel Popkin: senior research staff, National Bureau of Economic Research, Jan. 1977.

Arturo C. Porzecanski, Centro de Estudios Monetarios Latinoamericanos: economist, Morgan Guaranty Trust Co., New York, Feb. 1, 1977.

John Raisian, University of Washington: assistant professor, University of Houston, Sept. 1976.

Roy J. Ruffin, Carleton College: professor of economics, University of Houston, Jan. 1977.

William J. Stull: associate professor, department of economics, School of Business Administration, Temple University.

Howard P. Tuckman: director, Center for the Study of Education and Tax Policy, Institute for Social Research, Florida State University, Sept. 15, 1976.

Leaves for Special Appointments

Kurt Hausafus, Oberlin College: Chicago Mercantile Exchange, 1976-77

James R. Prescott, Iowa State University: economic advisor to the Iranian government, Dec. 1, 1976

Douglas K. Pearce, University of Houston: visiting assistant professor, Université Paris XII, Jan. 1977.

David Segal, Oberlin College: Graduate School of Design, Harvard University, 1976 77.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style.

A Please use the following categories:

1 - Deaths
2 - Retirements
3 - Foreign Scholars (visiting the USA or Canada)
4 - Promotions
5 - Administrative Appointments
6 - New Appointments
7 - Leaves for Special Appointments (NOT Sabbaticals)
8 - Resignations
9 - Miscellaneous

B Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), and the date at which the change will occur

C Type each item on a separate 3x5 card and please do not send public relations releases

D The closing dates for each issue are as follows. *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

A Monetary Model of Exchange Market Pressure Applied to the Postwar Canadian Experience

By LANCE GIRTON AND DON ROPER*

The monetary approach to the balance of payments has received considerable attention.[1] However, most of the empirical studies employ models of a small country with fixed exchange rates. Without relying on the small-country assumption, a model is derived to explain both exchange rate movements and official intervention. The dependent variable, which we call exchange market pressure, provides a measure of the volume of intervention necessary to achieve any desired exchange rate target. The model is applied to the postwar Canadian experience.

The paper is organized into three sections. The first section contains a discussion of the monetary approach and the problem of determining the degree of independence of a country's monetary policy. In the second section a monetary model that holds for all exchange rate regimes is developed. Empirical estimation of Canadian exchange market pressure and an empirical measure of the autonomy that the Canadian authorities relinquish when pursuing a fixed exchange rate target are presented in the third section.

I. The Monetary Approach and the Independence Question

If the balance of payments is divided into more than two accounts—for example, the current, capital, and money accounts—then each account can be explained with a direct or an indirect approach. Using the demands and supplies for the kth item as a classification procedure for explaining the kth account constitutes a direct approach. An account can be explained indirectly by first explaining the other n-1 accounts and then adding the results. Given that economists are accustomed to explaining the current and capital accounts (and various subaccounts) directly, the argument for the monetary approach can be seen as an argument that the money account be given symmetric treatment.[2]

The argument for symmetry also implies that the traditional view of official intervention (necessary to maintain fixed rates) as accommodating or financing current and capital account transactions should be abandoned.[3] The monetary approach continues to regard the quantity of intervention necessary to achieve a fixed rate target as endogenous, but it shifts the explanation to the monetary equilibrium condition.

By Walras' Law the net excess supply of goods and securities by residents[4] of a coun-

*Economist, Board of Governors of the Federal Reserve System, and associate professor of economics, University of Utah, respectively. We wish to acknowledge the technical assistance of Patrick Decker and Coralia Flaifel. We received helpful comments from George Borts, Michael Connolly, Michael Dooley, Charles Freedman, Jay Levin, Hyunchul Shin, and Jerome Stein. Extended discussions with Dale Henderson and Bennett McCallum were particularly useful. The views expressed in the paper are solely our own and do not necessarily represent the views of anyone else in the Federal Reserve System.

[1]See Bijan Aghevli and Mohsin Khan, Borts and James Hanson, Connolly and Dean Taylor, Rudiger Dornbusch, Jacob Frenkel and Carlos Rodriguez, Hans Genberg, Harry Johnson, Ryutaro Komiya, Donald McCloskey and Richard Zecher, Norman Miller and Sherry Askin, Robert Mundell (1968, 1971), Michael Mussa, and Zecher. A survey of this literature is given by Marina Whitman.

[2]Surprisingly, the theoretical advantages of the direct over the indirect approach has not, to our knowledge, been demonstrated for any of the accounts. The direct approach may have the practical advantage of increasing the likelihood of correctly specifying the explanatory functions.

[3]The argument that the items placed "below the line" were accommodating other exchange market transactions was emphasized in the discussions over the appropriate definition of the balance of payments in the early 1960's as found, for instance, in the "Bernstein Report" (see the U.S. Budget Bureau).

[4]"Resident" here means holders of cash balances whose demand is influenced by domestic income. The analysis in this paper will assume that the income

try represent a net excess demand for money. Thus the traditional approach to the balance of payments that specified behavioral relationships for the trade and capital accounts contained an implicit monetary condition. However, the implicit monetary condition was not necessarily one that would have seemed reasonable if money supply and demand functions had been developed explicitly.

The empirical estimation of a monetary model of the balance of payments can be related to, but is not identical with, an empirical estimation of the degree of independence of monetary policy. One can explain an official settlements measure of the balance of payments without testing for independence, and one can test for independence without explaining the balance of payments. In this paper, however, both exercises are undertaken since the monetary approach provides a useful framework within which to estimate the degree of monetary autonomy.[5]

Monetary independence can be measured by the degree to which alterations in the domestic source of the monetary base lead to changes in the demand for domestic base and thereby the total quantity outstanding.[6] If the policy actions used to alter the domestic source of the base fail to influence the demand for base money, then the change in the domestic source will be offset by the official exchange market intervention necessary to achieve a fixed exchange rate target.[7]

II. The Model

The model developed here, and analyzed further in the Appendix, is a monetary model in the sense that it organizes the analysis around the demands and supplies of national monies.

Using an exponential specification of the demand-for-base function, the monetary equilibrium condition for any country i can be written as[8]

$$(1) \quad H_i = F_i + D_i = P_i Y_i^{\beta_i} \exp(-\alpha_i \rho_i)$$

where H_i = supply of base money issued by the central bank of country i
F_i = base money created against the purchase of foreign assets
D_i = base money created by domestic credit expansion
P_i = price level

variable for one country affects the demand for the liabilities of the central bank of that country only. Implications of relaxing this assumption have been developed in our referenced paper.

[5]The usefulness of the monetary framework for dealing with the question of monetary independence may explain some of the recent interest in the approach. The growth of financial capital movements in the 1960's made policymakers aware of the conflict between their domestic monetary objectives and the commitment to a fixed exchange rate. One of the early studies concerned with this problem was by Ruth Logue.

[6]The most obvious way the domestic monetary authorities can affect the demand for their liabilities is by changing reserve requirements. In this paper, the impact of reserve requirement changes is subsumed under the supply side by adjusting base money for reserve requirement changes. The independence question concerns the ability to affect this adjusted base.

[7]A more general analysis of the question of autonomy would require that the sources of the base be divided into directly controlled and uncontrolled parts The purpose would then be to determine the degree that the latter tended to offset the former. The fact that there may be important uncontrolled elements other than official settlements is highlighted by the Federal Reserve's experience in the early 1920's when, in an effort to increase earnings, they were led to the "discovery" that open market operations were a policy tool equivalent to discounting. According to Ralph Burgess, "... as fast as the Reserve Banks bought Government securities in the market, the member banks paid off more of their borrowings; and, as a result, earning assets and earnings of the Reserve Bank(s) remained unchanged," (p. 221). Realization that discounting was offsetting open market operations made the Fed aware that the two procedures for purchasing domestic assets were good substitutes in their effect on bank reserves We assume these other potential offsets to monetary policy are sufficiently under the control of the authorities that intervention in pursuit of an exchange rate target can be usefully isolated as the primary threat to monetary autonomy This assumption allows us to address the two separate problems—explaining the exchange market pressure and measuring the degree of independence—within the same theoretical framework.

[8]One can specify the demand for base as the product of a money multiplier and the demand for an aggregate (defined as all financial items that absorb base money) Although such a specification may be useful for some purposes, it will not be used here.

Y_i = real income
ρ_i = index of interest rates
β_i = income elasticity > 0
α_i = interest rate coefficient > 0

The division of H between its domestic, D, and foreign, F, sources is determined by

$$F_i(t) = \int_{-\infty}^{t} E_i(\tau)R_i'(\tau)d\tau \tag{2}$$

where $R_i(t)$ = stock of international reserves (primary assets) held by the authorities in country i
$R_i'(t)$ = time derivative of R_i denoting net purchases at time t
$E_i(t)$ = parity or i currency value of primary reserve assets at time t

As the formula notes, the country's parity (or price of foreign exchange in the case of foreign exchange reserves)[9] is important only at the time foreign assets are purchased. If, for instance, the monetary authority devalues its currency and acquires a capital gain on their stock of international reserves, F as defined by (2) is not affected. As a result of the capital gain the authorities may increase their liabilities outstanding. But any increase in base money related to the capital gain should be treated as an increase in D, not F, since the purpose of the model is to explain the quantity of base that the authorities are induced to create or destroy (and the autonomy they sacrifice) in order to stabilize the exchange rate.

Substituting the time derivative of (2), viz., $F_i' = E_iR_i'$, in the differentiated version of (1) and stating the results in percent changes yields

$$h_i = r_i + d_i = \pi_i + \beta_i y_i - \alpha_i \rho_i' \tag{3}$$

where[10] $h_i = H_i'/H_i$, $d_i = D_i'/H_i$, $\rho_i'(t) = d\rho_i/dt$, $\pi_i = P_i'/P_i$, $r_i = E_iR_i'/H_i$, $y_i = Y_i'/Y_i$

By deflating the rate of change of international reserves valued in domestic currency E_iR_i' by domestic base money H_i, a real measure of the balance of payments r_i is obtained. It is essential to convert the nominal measure of the official intervention into real terms to determine whether the balance of payments is large or small.

To examine the monetary interaction between countries, subtract the monetary equilibrium condition (3) for country j from the monetary equilibrium condition for country i:

$$r_i - r_j = -d_i + d_j + \beta_i y_i - \beta_j y_j + \pi_i - \pi_j - \alpha(\rho_i' - \rho_j') \tag{4}$$

where α_i and α_j have been assumed equal ($\alpha = \alpha_i = \alpha_j$). We introduce the further notation,

e_{ij} = rate of appreciation of currency i in terms of currency j
$\theta_{ij} = \pi_i - \pi_j + e_{ij}$ = differential inflation rate adjusted for exchange rate changes[11]
$\delta_{ij} = \rho_i' - \rho_j'$ = change in the uncovered interest differential

Equation (4) can be rewritten as

$$r_i - r_j + e_{ij} = -d_i + d_j + \beta_i y_i - \beta_j y_j + \theta_{ij} - \alpha\delta_{ij} \tag{5}$$

The way equation (5) is employed to explain the interaction between two countries depends on whether one of the countries is sufficiently "large" in the sense of being able to pursue an independent monetary policy. Consider first the case of two regions

[9]International reserves can be expressed to include foreign exchange in which case (2) can be written as

$$F_i = \int_{-\infty}^{t} E_iR_i' + \int_{-\infty}^{t} E_{ij}R_{ij}'$$

where E_i is the ith currency value of the primary asset and E_{ij} is the ith currency value of the foreign exchange

[10]Taking into account foreign exchange, the definition of r_i is $E_iR_i'/H_i + E_{ij}R_{ij}'/H_i$. Primes denote derivates with respect to (the implied argument) time.

[11]Except where relative purchasing power parity (i.e., $\theta = 0$) is adopted for expository convenience in the Appendix, PPP, in either its absolute or relative versions, is not assumed in the paper; θ has been introduced for notational convenience only.

or countries of comparable size;[12] for example, France and Germany. If the mark-franc rate were perfectly fixed ($e_{ij} = 0$), the left-hand side of equation (5) would represent the bilateral (real) balance of payments. If both countries refrained from intervention ($r_i = 0 = r_j$), the left-hand side would reduce to the percent change of the mark-franc rate.[13] If the monetary authorities of the two countries intervened without a commitment to a perfectly constant exchange rate, the composite variable $r_i - r_j + e_{ij}$, measures what we refer to as exchange market pressure.

We are interested in applying the equation to Canada and the United States. Since the United States has been a center or key-currency country, it has had the ability to force most and perhaps all the adjustment burden on those countries who have made efforts to stabilize their exchange rates.[14] This extreme asymmetry in the adjustment burden justifies (as will be explained below) the transference of the center country's balance of payments from the left- to the right-hand side of the equation. If the i subscripts are changed to c (for Canada) and the j subscripts are changed to u (for the United States), then equation (5) can be rewritten as[15]

$$(6) \quad r_c + e_c = -d_c + h_u + \beta_c y_c - \beta_u y_u + \theta_c - \alpha\delta_c$$

where r_u has been subsumed under h_u ($= d_u + r_u$).[16]

The center country's balance of payments r_u can be taken to the right-hand side and used as an independent variable if h_u is not influenced by the remaining expression, $r_c + e_c$. Since r_u reflects changes in *U.S.* international reserves, then any official Canadian intervention financed by purchases or sales of *U.S.* dollars to the U.S. Treasury shows up in both r_c and r_u.[17] If *U.S.* reserve flows are perfectly sterilized,[18] however, then h_u ($= d_u + r_u$) is unaffected by r_c.[19] The additional observation that h_u has been managed independently of e_c implies that h_u can be taken as an exogenous variable in the equation.

[12]From the criterion of monetary autonomy or the distribution of the adjustment burden, the proximate determinates of "size" are the relative magnitudes of base money markets and, especially, the abilities of the monetary authorities to sterilize. Other parameters usually regarded as defining relative country size, especially real national income or wealth, are important only if they allow the authorities to sterilize more or to the degree that they determine the size of a country's base relative to the total base money of those countries with fixed exchange rate targets. This point is demonstrated in the Appendix.

[13]In the case of a pure float, the model is similar in spirit to what Frenkel has referred to as a "monetary model of exchange rate determination." A similar model of the exchange rate is developed by Bluford Putnam and John Woodbury.

[14]Unless the monetary authority is using exchange market intervention as an instrument to further domestic economic goals, an exchange rate target typically comes at the expense of domestic goals. The problem of monetary autonomy arises as a result of a fixed exchange rate *target* of which a fixed exchange rate is only one example. It is the rigidity of the target (which may be moving) rather than the rigidity of the rate that matters. A fixed exchange rate target means that the authorities are unwilling to trade this target off against other targets. A model is derived in the Appendix in which the monetary authority with a targeted growth path for the exchange rate loses all control over their domestic money growth rate.

[15]Equation (6) can also be derived from a multi-country model as shown in the Appendix.

[16]When the *U.S.* dollar is used as a numeraire, the second subscript u has been dropped in order to simplify the notation. In symbols, $e_c = e_{cu}$, $\theta_c = \theta_{cu}$ and $\delta_c = \delta_{cu}$.

[17]The h_u term, like the growth of base money for other countries, can be found by adding the domestic and foreign sources, $h_u = d_u + r_u$. But the international transactions that affect the supply of *U.S.* base money include official sales and purchases (not to be confused with allocations) of primary assets and not changes in foreign official holdings of *U.S.* dollar assets (except those that absorb base money). The r_u term, therefore, does not represent the *U.S.* official settlements or other measures of the *U.S.* balance of payments that have been traditionally employed. For noncenter countries, of course, r_i does represent the (deflated) official settlements balance of payments.

[18]Sterilization can be represented by breaking d_i into two components, viz., $d_i = d_i^o - \lambda_i r_i$ where λ_i is a sterilization coefficient ranging between unity (for complete sterilization) and a negative number (representing a reinforcement of the balance of payments necessary to play by the rules of the gold standard). Since d_i^o is exogenous or independent of r_i, then $h_u = d_u^o - \lambda_u r_u + r_u$ is independent of r_u and r_c if $\lambda_u = 1$.

[19]The notion of a "dollar standard"—a phrase developed in the 1920's when the United States abandoned the rules of the gold standard—is usually taken to mean that other countries adjust to the United States. In an important sense, noncenter countries sterilize *for* the center country and force more of the

The fact that the *U.S.* monetary policy has been insulated in the postwar monetary system (such that h_u in equation (6) can be taken as independent of r_c and e_c) allows further flexibility in the way the model is specified. If *U.S.* monetary policy had not been independent of the balance of payments, the monetary interaction between the United States and other countries would have been through both the supply and demand sides of base money markets. With h_u unaffected by exchange market intervention, the link between the United States and the rest of the world is only through the demand side—substitution between securities and commodities. The link is a recursive one that goes from *U.S.* prices and interest rates, to c prices and interest rates, to the demand for c's base, to the induced supply of base, r_c.

The absence of the supply link (the fact that r_c does not feed back on h_u) means that h_u need *not* appear in the equation. Equation (6) can be written to capture the linkages on the demand side by including *U.S.* prices and interest rates on the right-hand side of (6) while excluding h_u.[20] Since we want to determine the influence of *U.S.* monetary policy on Canadian exchange market pressure, however, it is useful to have a one-variable index of *U.S.* monetary policy rather than the two variables (interest rates and prices) over which the *U.S.* authorities have less control.

III. Empirical Investigation

Equation (6) differs from equations in other monetary[21] models of the balance of payments in two ways. First, the dependent variable is exchange market pressure, defined as the sum $r + e$, rather than the balance of payments, per se. If the value of the dependent variable $r + e$ is unaffected by its composition (as will be subsequently measured as e/r), then the exchange market pressure is independent of whether the authorities absorb the pressure in their reserves or in their rate. Second, the equation takes account of the fact that, as far as foreign exchange market pressure is concerned, a country's monetary policy can be judged tight or easy only by reference to what is happening in the rest of the world. Consequently, the country's external position is related to foreign monetary conditions. The supply and demand for *U.S.* money are used to represent world monetary conditions.

The purpose of this section is both to estimate the monetary equation (6) of exchange market pressure and to measure the degree to which the central bank in an open economy can pursue an independent monetary policy. Estimation of equation (6) in its present form would serve the first purpose of explaining exchange market pressure, but it would not provide a measure of monetary independence. If equation (6) were estimated, a minus-one coefficient in front of d_c would be expected regardless of whether there was multicollinearity between d_c and δ_c or θ_c. The crucial issue in determining the degree that a fixed exchange rate target undermines monetary autonomy is whether the authorities can make their interest rates and prices diverge from *U.S.* interest rates and prices by the use of monetary policy. In symbols, the measure of independence is the

adjustment burden on themselves when they acquire or lose dollar assets rather than outside reserve assets. Further discussion of the sense in which other countries sterilize for the United States and a discussion of their incentives for forcing more of the adjustment burden on themselves is found in Girton and Henderson, and in Roper.

[20]Since monetary equilibrium requires $h_u = \pi_u + \beta_u y_u - \alpha\rho'_u$, $\pi_u - \alpha\rho'_u$ can be substituted for $h_u - \beta_u y_u$ in equation (6) to obtain

$$r_c + e_c = -d_c + \beta_c y_c + \pi_u - \alpha\rho'_u - \alpha\delta_c + \theta_c$$

[21]Two possible reasons for referring to equation (6) as a "monetary" model should be sharply distinguished. Suppose that d_c and h_u were very stable over time and that y_c and y_u were subject to large fluctuations due, say, to earthquakes. If the model were named for the independent variables with the highest variance and greatest potential explanatory power, it would be an "earthquake" model. If it is named for the fact that monetary equilibrium conditions are being used as an organizing framework for explaining $r + e$, it would be a monetary model. Following this second line of reasoning, equation (6) would still therefore be a monetary model even if there were no variance, and no explanatory power, in the d_c and h_u variables.

degree to which δ_c and θ_c depend on the Canadian control variable d_c.

To develop an alternative to (6) that will allow one to measure the independence of monetary policy, suppose that δ_c and θ_c are determined by the reduced form relations:

$$(7) \quad \delta_c = \delta(d_c, h_u, X) \qquad \theta_c = \theta(d_c, h_u, X)$$

$$\text{where} \quad \delta_1 = \frac{\partial \delta}{\partial d_c} \leq 0 \qquad \delta_2 = \frac{\partial \delta}{\partial h_u} \geq 0$$

$$\theta_1 = \frac{\partial \theta}{\partial d_c} \geq 0 \qquad \theta_2 = \frac{\partial \theta}{\partial h_u} \leq 0$$

and X is the set of other variables that influence δ_c and θ_c.

If Canadian and *U.S.* securities and goods are not perfect substitutes then anything that affects the supplies and demands for Canadian securities and goods relative to *U.S.* securities and goods will be in the set of X variables. Changes in monetary conditions outside the United States and Canada should be included only if they affect Canadian and *U.S.* securities and goods markets differentially. Since Canadian and *U.S.* real incomes might have differential effects on Canadian and *U.S.* prices and interest rates, the estimated income coefficients could be affected by any imperfect substitutability of Canadian and *U.S.* goods and securities.

Assuming the expressions in (7) are linear, they can be substituted into equation (6) to obtain

$$\begin{aligned}(8) \quad r_c + e_c &= -(1 + \alpha\delta_1 - \theta_1)d_c \\ &\quad + (1 - \alpha\delta_2 + \theta_2)h_u \\ &\quad + \beta_c y_c - \beta_u y_u \\ &\quad + (\theta_x - \alpha\delta_x)X \\ &= -\phi_c d_c + \phi_u h_u + \beta_c y_c \\ &\quad - \beta_u y_u + (\theta_x - \alpha\delta_x)X\end{aligned}$$

$$\text{where } \phi_c = 1 + \alpha\delta_1 - \theta_1$$
$$\phi_u = 1 - \alpha\delta_2 + \theta_2$$

To the extent that d_c affects δ_c or θ_c, the estimated value of ϕ_c should be less than unity. That is, the Canadian reserve loss or exchange rate depreciation associated with an expansionary monetary policy will be mitigated if the policy lowers Canadian interest rates relative to *U.S.* rates, or raises Canadian prices relative to *U.S.* prices. Assuming that ϕ_c is the same under floating as under fixed rates, it is legitimate to use the data generated under floating rates to estimate the degree of independence the authorities lose by the adoption of a fixed rate target.

The form of the equation to be estimated is

$$(9) \quad r_c + e_c = -\phi_c d_c + \phi_u h_u + \beta_c y_c - \beta_u y_u + v$$

where v is a random term. The exclusion of the X variables will not bias the estimated coefficients if the X variables are uncorrelated with the right-hand variables of equation (9).

To justify an *OLS* estimation of (9), argument must be provided for the recursiveness of the relationship. Each of the terms on the right-hand side of (9) will be discussed to determine whether they can be regarded as independent of the random term v.

Consider first the *U.S.* monetary policy and income variables. In Section II it was argued that *U.S.* monetary aggregates have been independent of r_c and this independence also eliminates the only obvious channel through which r_c might have influenced *U.S.* income. It is also unlikely that e_c would affect y_u or h_u. Consequently, y_u and h_u can be treated as independent of the error term regardless of whether the Canadian authorities absorb market pressure in their reserves or exchange rate.

If other *U.S.* monetary aggregates are independent of $r_c + e_c$, then the choice of which *U.S.* monetary aggregate to use depends on which one is the best indicator of *U.S.* monetary conditions. Since other *U.S* monetary aggregates might be good indicators of *U.S.* monetary conditions, we report results using two additional aggregates, money narrowly defined ($M1_u$) and a broader measure ($M2_u$). Although this freedom of choice exists for the *U.S.* aggregate, it does not exist for any country whose supply of base money is influenced by exchange market intervention.

The rate of growth of Canadian real in-

come should be independent of r_c and e_c to the extent that y_c is based on past income changes. Since y_c is measured with a distributed lag in the regressions, all of the past values of y_c should be independent of the error term, but the possibility of simultaneous equations bias from the current change in real income cannot be ruled out.[22]

Potentially, the most important simultaneity problem occurs in the estimation of ϕ_c under fixed exchange rates.[23] If the Canadian authorities try to sterilize reserve flows, $\hat{\phi}_c$ will be biased regardless of the success of their sterilization policy. It is generally difficult to determine the direction and amount of bias for the estimated coefficients in a multiple regression. It is useful, however, to consider the bias under the simplifying assumption that h_u, y_u, and y_c are uncorrelated with d_c when the exchange rate is fixed. In this case $\hat{\phi}_c$ can be shown to have an asymptotic bias of $1/\lambda_c$ (or zero) as the variance of d_c relative to the variance of v approaches zero (or infinity).[24] Under a floating rate regime, however, any bias of $\hat{\phi}_c$ is in the opposite direction. The fact that the Canadian dollar remained around unity vis-à-vis the *U.S.* dollar suggests that the authorities responded to a positive e_c by making d_c larger than otherwise.[25] If this positive effect were represented in a linear fashion by Λ_c (>0), then ϕ_c could be biased toward $1/\Lambda_c$.

Equation (9) is estimated using annual data for the period 1952 through 1974. During this period, the Canadian dollar floated from 1952 to 1962 and after June of 1970, and was fixed in value to the *U.S.* dollar in the intervening years. A wide range of domestic policies were pursued, and there were various agreements concerning Canadian-*U.S.* economic relations.[26]

Two kinds of adjustments were made to the data. First, the series on the stock of Canadian international reserves was adjusted to exclude gold revaluation gains and *SDR* allocations.[27] Second, the Canadian base money figures were adjusted for the reserve requirement changes that occurred in 1952 and 1967.

In the regressions reported in Table 1,

[22]One might expect that when the Canadian dollar appreciates ($e_c > 0$), the export industries and the import competing sector might have to contract and this could dampen current real output. If y_c is independent of the other explanatory variables, this would bias β_c downwards. A downward bias is consistent with some of the results that show the estimated value of β_c to be slightly lower for current values of y_c than for lagged values of y_c when the lags were unconstrained. If the lag between the other part of the dependent variable r_c and current output were sufficiently long, then the impact of e_c on y_c would be the only channel for simultaneity bias in the estimate of β_c.

[23]To our knowledge, the study that has come closest to separating the impact of r on d (sterilization) from the impact of d on r (the offset) is by Michael Porter. He estimated a monthly model of the German capital account in which his d variable was primarily a reflection of changes in reserve requirements. Since the Bundesbank alters reserve requirements at the beginning of each month, then that month's capital flow could be taken as depending on the reserve requirement changes at the first of the month. There would be no feedback from the capital movement to the policy variable unless the authorities were anticipating the future capital flows.

[24]If the variance of the random term v is labeled σ_v and the variance of d_c^o is σ^*, then the asymptotic bias is given by the formula $plim(\hat{\phi} - \phi) = (1 - \lambda\phi)\lambda/(\sigma^*/\sigma_v + \lambda^2)$ where the c subscripts have been omitted for convenience. If $\sigma^*/\sigma_v = 0$, $plim(\hat{\phi}) = 1/\lambda$. If σ^*/σ_v approaches infinity, the bias approaches zero.

Pentti Kouri and Porter derive a formula for the asymptotic bias that is correct (if they assume their X_1 variable is independent of their *NDA* variable), but their verbal discussion of the bias is misleading. They assert (pp. 453–54) that their estimated coefficient (corresponding to our $\hat{\phi}_c$) is biased towards unity rather than towards the reciprocal of the sterilization coefficient. The bias will be towards unity only if the sterilization coefficient is unity. If the authorities only attempt to partially sterilize, then the bias will be towards a number greater than one.

[25]In Figure 1, the variance of the dependent variable $r_c + e_c$ appears greater for periods of fixed rates than for periods of floating rates. This is consistent with the argument that the authorities responded asymmetrically to a given pressure in the exchange market depending on the exchange rate regime. To the extent the authorities attempted to sterilize reserve flows, exchange market pressure was increased, and to the extent they attempted to dampen exchange rate movements during floating rate periods, exchange market pressure was reduced.

[26]See Robert Dunn for a discussion of the domestic policies and their objectives and the various Canadian-*U.S.* economic agreements.

[27]In principle, the international reserve figures should also be adjusted for several other types of nonmarket transactions. We did not do this because of data limitations.

TABLE 1

Aggregate Used	Constant	d_c	Coefficients of h_u	y_c	y_u	ρ	R^2	S.E.	D.W.
$M2_u$	$-.04$	$-.96$	1.14	2.80	-2.84	.22	.92	.024	1.80
	(1.08)	(12.74)	(4.86)	(3.01)	(3.59)	(1.07)			
$M1_u$	$-.03$	$-.96$	1.74	2.54	-2.51	$-.06$	.95	.020	2.11
	(1.38)	(16.03)	(8.37)	(3.97)	(4.83)	(.28)			
H_u	$-.03$	$-.97$	1.61	2.63	-2.62	.06	.96	.017	2.29
	(1.43)	(18.53)	(9.09)	(4.46)	(5.35)	(.30)			

Note: The dependent variable in all the regressions is $r_c + e_c$, the measure of exchange market pressure. The values in parentheses below the coefficients indicate t-ratios for the corresponding estimates. The S.E. and D.W. denote, respectively, the standard error of the regression and the Durbin-Watson coefficient. R^2 is the coefficient of determination adjusted for degrees of freedom. The income coefficients are the sum of four-year distributed lags of real GNP for each country. A second-degree Almon polynomial was used with the far tail tied to zero. All equations were run using the Cochrane-Orcutt technique to adjust for serial correlation; ρ is the estimated value of the first-order autoregression coefficients.

equation (9) is estimated using the percent change in three alternative U.S. monetary aggregates representing U.S. monetary conditions. The first regression uses $M2_u$, the second uses money narrowly defined ($M1_u$), and the third uses U.S. base money (H_u). The estimated coefficients have the correct sign and are significant at the 5 percent confidence level.[28] The explanatory power of the model is illustrated in Figure 1 where the actual and predicted values of the dependent variable are plotted. The predicted values are those coming from the regression in which a broad U.S. monetary aggregate is used.

To interpret the implication of the estimated value of the ϕ_c coefficient, it is useful to consider the degree that Canadian monetary independence would be curtailed if ϕ_c were .95. Suppose that in a particular year the expression $\phi_u h_u + \beta_c y_c - \beta_u y_u + v$ is zero so that equation (9) reduces to $r_c + e_c = -.95 d_c$. Suppose the Canadian authorities initially attempt to increase their money growth rate by 10 percent by setting $d_c = .10$. Then, they must be prepared to either allow their currency to depreciate by 9.5 percent during the year ($e_c = -.95(.10) = -9.5$ percent) or lose reserves at a rate equal to 9.5 percent of their base ($r_c = -.95(.10) = -9.5$ percent) or some combination.

Alternatively, suppose the authorities purchase domestic assets in sufficient quantity to successfully raise their base from C$10 to C$11 billion. Then a .95 value for ϕ_c implies that C$9.5 billion worth of foreign reserves would be required by the Canadian authorities to support their rate. The estimated coefficient on the domestic source of Canadian base money (d_c), of $-.96$ or $-.97$, supports the view that the Canadian monetary authorities, when under a fixed exchange rate regime, have little scope of pursuing an independent monetary policy.

An alternative way of expressing equation (9) that highlights the implications for monetary independence is

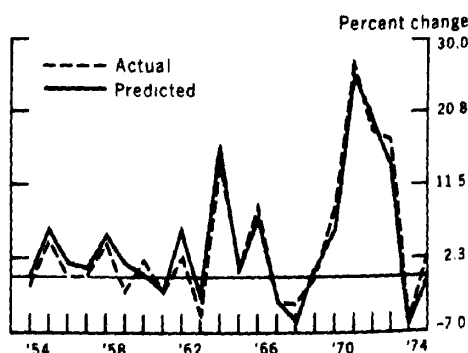



FIGURE 1. EXCHANGE MARKET PRESSURE ($r_c + e_c$)

[28]The model has no implication for the sign or significance of the constant term. The constant term is not significant.

$$(9') \quad h_c + e_c = I_c d_c + \phi_u m_u + \beta_c y_c - \beta_u y_u + v$$

Equation (9′) is found by adding d_c to both sides of equation (9) such that $I_c = 1 - \phi_c$. Since h_c is on the left-hand side of (9′), it is clear that the estimate of I_c is a measure of the degree to which d_c influences h_c when the Canadian authorities keep $e_c = 0$. The coefficients obtained from estimating (9′) would be identical with those obtained from the estimation of (9), except that $\hat{I}_c = 1 - \hat{\phi}_c$. Our estimated value of $\hat{\phi}_c$ above .95 implies a value for $\hat{I}_c$ below .05. We use (9) instead of (9′) because we are interested in explaining Canadian foreign exchange market pressure as well as estimating the degree of Canadian monetary independence.

The measure of exchange market pressure used as the dependent variable in the regressions can be split into the two components, changes in official reserves, and changes in exchange rates. The assumptions used imply that the total of these two components is not sensitive to the composition. In order to test for the sensitivity of the measure of exchange market pressure to its composition (whether the authorities absorb pressure in international reserves or in the exchange rate), the equations were reestimated with the ratio $Q = e_c/r_c$ entered as a separate explanatory variable.[29] The results were the same for all three equations. For that reason, only the $M2_u$ equation is reported:

$$(10) \quad r_c + e_c = \underset{(1.03)}{-.04} \underset{(12.21)}{-.94d_c} + \underset{(4.81)}{1.12h_u} + \underset{(3.41)}{2.90y_c} - \underset{(4.15)}{3.03y_u} + \underset{(1.08)}{.001Q}$$

$$\rho = \underset{(.36)}{.08}, \quad R^2 = .92$$

$$S.E. = 2.025, \quad D.W. = 2.06$$

[29] Another way of testing for the sensitivity of the dependent variable to its composition would be to split the dependent variable into its components and put one of the components on the right-hand side as an exogenous variable. But doing this would introduce a simultaneity problem since the authorities often react to exchange market pressure by both intervening and allowing the exchange rate to move.

The coefficient on Q is not significant and other coefficients are left essentially unchanged indicating that the explained value of exchange market pressure is not sensitive to its composition. This implies the dependent variable $r + e$ can be appropriately used to determine the volume of intervention necessary to achieve various exchange rate targets.

APPENDIX

In the text, monetary conditions in the United States and Canada were used to develop an equation for estimating a measure of exchange market pressure for Canada. Here we show that the results obtained from focusing on only the two countries are consistent with the limiting case in an explicit multicountry framework. Also, several propositions that were asserted in the text are proved.

The flow monetary equilibrium condition for country i is

$$(A1) \quad h_i = r_i + d_i = \pi_i - \alpha\rho_i' + \beta_i y_i$$

The policy reaction function for country i is expressed as

$$(A2) \quad d_i = d_i^o - \lambda_i r_i$$

or

$$h_i = d_i^o + (1 - \lambda_i) r_i$$

Using the policy reaction function (A2) to substitute for d_i in (A1) and solving for r_i yields

$$(A3) \quad r_i = \frac{\pi_i - \alpha\rho_i' + \beta_i y_i - d_i^o}{(1 - \lambda_i)}$$

The world demand for international reserves is assumed to be equal to the supply. This world reserve equilibrium condition can be expressed as

$$(A4) \quad \Sigma s_i r_i = r_w$$

where $s_i = H_i/E_i H_w$, $H_w = \Sigma H_i/E_i$

and r_w is the rate of change in the supply of international reserves as a proportion of world base money (H_w).

Substituting the demand for international

reserves (A3) into the global reserve equilibrium condition (A4) yields

(A5)
$$\sum \frac{s_i}{(1-\lambda_i)}(\pi_i - \alpha\rho_i' + \beta_i y_i - d_i^o) = r_w$$

Defining θ_i and δ_i analogous to the definitions of θ_c and δ_c, i.e., for country i with respect to the United States, substituting the θ_i and δ_i in (A5), and solving for $n = \pi_u - \alpha\rho_u'$ yields

$$\text{(A6)} \quad n = \frac{r_w}{\Sigma s_i/(1-\lambda_i)} - \Sigma w_i(\beta_i y_i - e_i + \theta_i - \alpha\delta_i - d_i^o)$$

where $w_i = \dfrac{s_i/(1-\lambda_i)}{\Sigma s_i/(1-\lambda_i)}$

An expression for h_c in terms of the policy parameters and the other exogenous variables is found by substituting θ_c, δ_c, and the expression for n into (A1):

$$\text{(A7)} \quad h_c = \frac{r_w}{\Sigma s_i/(1-\lambda_i)} - \Sigma w_i(\beta_i y_i - e_i + \theta_i - \alpha\delta_i - d_i^o) + (\beta_c y_c - e_c + \theta_c - \alpha\delta_c - e_c)$$

The expression for r_c is obtained by substituting the expression for h_c into the policy reaction function (A2):

$$\text{(A8)} \quad r_c = \frac{1}{1-\lambda_c}\left[\frac{r_w}{\Sigma s_i/(1-\lambda_i)} - \Sigma w_i(\beta_i y_i - e_i + \theta_i - \alpha\delta_i - d_i^o) + (\beta_c y_c - e_c + \theta_c - \alpha\delta_c - d_c^o)\right]$$

Assuming the other right-hand side variables in (A7) are independent of d_c^o, i.e., the e are fixed and the θ and δ are independent of d_c^o, and that $\partial\lambda_i/\partial\lambda_j = 0$, for $i \neq j$, then

$$\text{(A9)} \quad \frac{\partial h_c}{\partial d_c^o} = w_c$$

$$\frac{\partial^2 h_c}{\partial d_c^o \partial\lambda_c} = w_c \sum_{i \neq c} w_i \geq 0$$

$$\frac{\partial^2 h_c}{\partial d_c^o \partial\lambda_u} = \frac{-w_c w_u}{1-\lambda_u} \leq 0$$

In a highly integrated world, with fixed exchange rate targets, $(0 \leq w_c \leq 1)$ determines the power of a country c, through domestic monetary operations, to influence its own and the world's rate of monetary growth.[30] The second and third expressions in (A9) show the change in the degree of influence over world monetary conditions country c has with changes in its own and others sterilization behavior. By inspection of (A9) it can be seen that, assuming no country completely sterilizes, w_c increases with increases in λ_c and decreases with increases in $\lambda_u (u \neq c)$. In the limit where country u completely sterilizes ($\lambda_u = 1$), then $w_c = 0$, for all $c \neq u$, and $w_u = 1$.[31]

When $\lambda_u = 1$ and $\lambda_c = 0$, the expressions for h_c and r_c reduce to

$$\text{(A10)} \quad h_c = h_u + \beta_c y_c - \beta_u y_u - e_c + \theta_c - \alpha\delta_c$$

and

$$\text{(A11)} \quad r_c + e_c = -d_c^o + h_u + \beta_c y_c - \beta_u y_u + \theta_c - \alpha\delta_c$$

where it is recognized that by definition e_u, θ_u, $\delta_u = 0$, and that when $\lambda_u = 1$, then $\Sigma s_i/(1-\lambda_i) = 0$ and $h_u = d_u^o$.

Data Appendix

Canadian International Reserve figures were obtained by adjusting "Canadian Official International Reserves-Total" (series B3800 in the *Bank of Canada Review* (*BCR*) for *SDR* allocations and gold revaluation profits. These adjustments take out the effect of nonmarket transactions on the reserve stock figures. There are other types of nonmarket transactions that we have not adjusted for because of data difficulties. Nonmarket transactions in reserve assets will tend to bias the coefficient on domestic assets toward negative one. The reserve series used consists

[30]The w, which depend on the s and λ, are the relevant measure of country size in the model used in this paper. It should be noted that this measure of size depends on the assumption that domestic open market operations should be scaled by the stock of base money.

[31]If more than one country completely sterilizes, the model is overdetermined.

of end-of-month stock figures. The r_c was calculated by first differencing the average of the end-of-month reserve figures and dividing by the product of lagged values of the exchange rate and Canadian base money.

Canadian Monetary Base was obtained by adding "Total coin outside banks" (*BCR* series B2003), "Total notes in circulation" (*BCR* series B51), and chartered bank deposits at the Bank of Canada (*BCR* series B55). Annual averages of Wednesday figures are used.

Domestic Assets Held by Canadian Monetary Authorities were obtained by subtracting Canadian International Reserves, in Canadian dollar terms, from the Canadian monetary base.

Exchange Rate (*U.S.* dollar price of Canadian dollars) is the annual average of noon buying rates in New York.

U.S. Monetary Base is the annual average of the weekly figures put out by the St. Louis Federal Reserve Bank.

U.S. and Canadian Real GNP figures are annual averages of quarterly figures put out by the U.S. Office of Business Economics and Statistics Canada.

REFERENCES

B. B. Aghevli and M. S. Khan, "The Monetary Approach to Balance of Payments Determination: An Empirical Test," Int. Monet. Fund DM/74/113, Nov. 1974.

V. Argy and P. Kouri, "Sterilization Policies and the Volatility in International Reserves," in Robert Z. Aliber, ed., *National Monetary Policies and the International Financial System*, Chicago 1974.

G. Borts and J. Hanson, "The Monetary Approach to the Balance of Payments," in Jere Behrman, ed., *Short-Run Macroeconomic Policy in Latin America*, forthcoming.

R. Burgess, "Reflections on the Early Development of Open Market Policy," *Fed. Reserve Bank of New York Mon. Rev.*, Nov. 1964, *46*, 219–26.

M. Connolly and D. Taylor, "Testing the Monetary Approach to Devaluation in Developing Countries," *J. Polit. Econ.*, Aug. 1976, *84*, 849–60.

R. Dornbusch, "Devaluations, Money, and Nontraded Goods," *Amer. Econ. Rev.*, Dec. 1973, *63*, 871–80.

Robert M. Dunn, *Canada's Experience with Fixed and Flexible Exchange Rates in a North American Capital Market*, Washington 1971.

J. A. Frenkel, "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scand. J. Econ.*, May 1976, *78*, 200–25.

——— **and C. A. Rodriguez**, "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach," *Amer. Econ. Rev.*, Sept. 1975, *65*, 674–88.

H. A. Genberg, "Aspects of the Monetary Approach to Balance of Payments Theory: An Empirical Study of Sweden," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

L. Girton and D. Henderson, "Financial Capital Movements and Central Bank Behavior in a Two-Country, Short-run Portfolio Balance Model," *J. Monet. Econ.*, Jan. 1976, *2*, 33–62.

——— **and D. Roper**, "Theory and Implications of Currency Substitution," Int. Fin. disc. pap. no. 86, Fed. Res. Board, 1976.

H. G. Johnson, "The Monetary Approach to Balance of Payments Theory," *J. Finance. Quant. Anal.*, Mar. 1972, *7*, 1555–72.

P. Komiya, "Economic Growth and the Balance of Payments: A Monetary Approach," *J. Polit. Econ.*, Jan./Feb. 1969, *77*, 35–48.

P. Kouri and M. G. Porter, "International Capital Flows and Portfolio Equilibrium," *J. Polit. Econ.*, May/June 1974, *82*, 443–67.

R. Logue, "Imported Inflation and the International Adjustment Problem," Staff Econ. stud. no. 55, Fed. Res. Board, 1969.

D. McCloskey and R. Zecher, "How the Gold Standard Worked: 1880–1913," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

N. C. Miller and S. S. Askin, "The Balance of Payments and Monetary Autonomy in Brazil and Chile," unpublished paper, 1975.

Feliks Mlynarski, *Gold and Central Banks*, London 1929.

Robert A. Mundell, *International Economics*, New York 1968.

———, *Monetary Theory*, Pacific Palisades 1971.

M. Mussa, "Tariffs and the Balance of Payments: A Monetary Approach," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

M. G. Porter, "Capital Flows as an Offset to Monetary Policy: The German Experience," *Int. Monet. Fund Staff Pap.*, July 1972, *19*, 395–424.

B. H. Putnam and J. R. Woodbury, "Exchange Rate Stability and Monetary Policy: A Case Study," Fed. Res. Bank New York, unpublished paper 1976.

D. Roper, "Implications of the Gold-Exchange Standard for Balance of Payments Adjustment," *Economica Internazionale*, Aug./Nov. 1973, *26*, 3–22.

M. Whitman, "Global Monetarism and the Monetary Approach to the Balance of Payments," *Brookings Papers*, Washington 1975, *3*, 491–555.

R. Zecher, "Monetary Equilibrium and International Reserve Flows in Australia," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

Bank of Canada, *Bank of Canada Review (BCR)*, Ottawa, various issues.

Board of Governors of the Federal Reserve System, *Foreign Exchange Rates*, H.10 Release, Washington, various issues.

Statistics Canada, Canadian Statistical Review, Ottawa, various issues.

U.S. Budget Bureau, *Balance of Payments Statistics of the U.S.: A Review and Appraisal*, Washington 1965.

U.S. Office of Business Economics, *Surv. Curr. Bus.*, Washington, various issues.

Hedonic Wage Equations and Psychic Wages in the Returns to Schooling

By ROBERT E.B. LUCAS*

Since the time when Adam Smith first observed that public hangmen received higher wage rates in compensation for their obnoxious task, the notion of equalizing wage differentials has maintained its status as a received doctrine, despite a dearth of systematic empirical work in the interim. This neglect can partially be attributed no doubt to the difficulties of interpreting such catch-all concepts as reported job satisfaction, but the discovery of an unusually rich and hitherto unexplored source of data renders possible a "new approach" to the topic in this paper.

These data, collected for the *Dictionary of Occupational Titles*, describe in terms of occupational attributes the nature of the work task involved in some 14,000 jobs. For example: the physical work environment is described by indicating the presence of toxic conditions, extreme temperatures, and hazard; whether the job is highly repetitive, or is supervisory are recorded; and the levels of such abilities as strength and general educational development "required" for task execution are estimated. The principal objective here is to discover how individuals' wages vary, *ceteris paribus*, with such indicators of the quality of working life, by inserting these job characteristic variables into a wage equation that also embraces personal data.

It is apparent that such a relationship truly belongs to the general class of hedonic price functions, yet, in contrast to existing hedonic equations, this wage equation embodies two quite distinct sets of characteristics: one describing people, the other describing their jobs.[1] The essential cause of this contrast is that the market for jobs does differ in a fundamental sense from the consumer goods markets as commonly conceived, for entrepreneurs are not indifferent to the identity of workers to whom they "sell" jobs, as is commonly supposed in the sale of consumer goods. Section I therefore augments the existing theory of hedonic prices for consumer goods by briefly considering the problems of choice on behalf of fully cognizant employees and employers facing parametric wages, in a scenario where both work and workers vary in quality as depicted by their separate characteristics.[2] Market clearing, in the sense of matching choices, is shown to generate a hedonic wage equation of the type to be estimated in Section II, being a reduced form outcome of a market wage-setting process.

Section II describes the *U.S.* cross-sectional data file compiled for this study, and presents the hedonic wage equations estimated. Section III illustrates one application of the estimated hedonic coefficients, namely in computing the direction and magnitude of bias resulting from estimating the rates of return to schooling from money wage data alone, when standard theory suggests that gains in the monetary equivalent of psychic wages ought also to be included.

I. Pairing of Workers and Jobs

A. *Workers' Choices*

The basic postulate in this choice model is that each worker cares both about the

*Boston University. I am grateful to the following persons for comments on earlier drafts: R. W. Clower, F. M. Fisher, R. E. Hall, G. Hanoch, J. G. Riley, F. R Welch, and the referee. The U.S. Department of Labor provided financial support under contract No. J-9-M-5-0042.

[1]Finis Welch and Zvi Griliches both noted the connection between equations relating wages to personal characteristics alone and the hedonic price literature.

[2]In particular, this is an extension to the labor-job market context of the hedonic models in Sherwin Rosen (1974) and the author (1975).

monetary rewards from his work and about the "quality" of his working life as described by some vector of attributes of the job performed, written here as vector Z_i for job i. (For a summary of notation adopted in this model, see the Appendix.) Of course, attitudes to pecuniary and nonpecuniary rewards vary from person to person, but it may be supposed that tastes are at least partially conditioned by some vector of measured personal characteristics of the worker (vector G^α for person α).[3] Two basic reasons may be given for such a postulate here:

a) A presumption of dissatisfaction resulting from performing a task that involves the use of more, or less, of an ability than is possessed by the person.[4]

b) Tastes are not determined in heaven; they are, at least partially, formed through environmental experiences, and some elements of the vector of personal characteristics may be viewed as factors that condition the probability of having had certain experiences.

The study of discrete choice in this context differs from existing work on consumer goods only in that prices cannot be assumed uniform. Not everybody is offered the same wage rate for any given job, so the wage rate offered to worker α for occupation i must generally be written as ω_i^α. In addition, not every job is a member of any one person's job choice set, yet at least conceptually, one can imagine such excluded jobs as being available but only at a wage rate which removes the job from the set of feasible solutions.

Each worker's utility may now generally

FIGURE 1

be written as some function:[5]

$$u^\alpha = u(\omega_i^\alpha, Z_i, G^\alpha, \epsilon^\alpha) \quad (1)$$

where ϵ^α represents the idiosyncratic element in α's taste not attributable to G^α, and not available to the empirical observer.

A choice problem for a group of workers with common wage offers for all relevant jobs, assuming a continuum of such jobs, is illustrated in Figure 1 for one job characteristic which is distasteful to all.[6] Workers choose from a set of jobs bounded by a frontier (comparable to the Lancastrian budget constraint) such as E_1E_1. Efficient job choosers from this set, possessing complete information and various tastes, distribute themselves along E_1E_1 at points where their indifference curves are tan-

[3]This general approach is developed by Daniel McFadden. For an application to occupational choice, see Michael Boskin.

[4]A person of high intelligence but low strength may enjoy being a chess player, but the same person may be repulsed by the prospect of playing football (at the same wage rate); another person, with a different comparative advantage, may reverse the ranking of these jobs. This idea is formalized by Jan Tinbergen who supposes utility to be determined by a quadratic loss function dependent upon discrepancies between job and personal attribute values, though Tinbergen's job characteristics are rather different from those above.

[5]Notice that leisure may be treated in the context of (1) in either of two ways: by viewing leisure as a "job" which the worker either does or does not perform; or by measuring hours of work as one job characteristic. On the former, see Kelvin Lancaster and on the latter Rosen (1969).

[6]The assumption of a job continuum is not essential to the following, though in its absence the Kuhn-Tucker conditions are only approximated.

gential to the frontier.[7] Of course, there exist at once many frontiers, such as E_1E_1 and E_2E_2, one for each group of workers with common wage offers, though each has a positive slope given a universal distaste for this job characteristic.

When a worker chooses a job, i, to maximize (1), the result is an equating of the marginal rates of substitution between wage rate and each job characteristic with the ratio of their shadow prices at the relevant point on the efficiency frontier. Solving these equalities provides a set of supply functions for each worker:[8]

$$\theta_i^\alpha = \theta(\Omega^\alpha, Z_i, G^\alpha, \epsilon^\alpha), \quad \text{all } i, \alpha \tag{2}$$

where $\theta_i^\alpha = 1$ if α selects occupation i, 0 otherwise;
Ω^α is the vector $[\omega_1^\alpha \ldots \omega_i^\alpha \ldots]$

B. *Firms' Choices*

It is fairly conventional to maintain that workers' productivities vary systematically according to some vector of personal characteristics, this being the tacit root of both human capital and signalling theory. However, in the present model, the characteristics of the job performed are also introduced into the function for net profit (π^α) generated by worker α. The arguments for introducing the latter are twofold: first, because productivity is likely to be influenced by the extent to which a person's abilities match those required in the execution of a job task;[9] secondly, that a person's willingness to supply effort while on the job probably depends upon the nature of the task involved.[10] Thus:

$$\pi^\alpha = \pi(\omega_i^\alpha, Z_i, G^\alpha, \xi^\alpha) \tag{3}$$

where ξ^α is the worker's idiosyncratic skill not reflected in G^α.[11]

It must be recognized that a firm makes, at any moment in time, two types of decision vis-à-vis a potential worker: whether to (continue to) hire, and to which job this person may best be assigned if hired. Conceptually, such decisions may be treated as if made simultaneously for every worker.

Again maximization leads to an efficiency frontier, ee, as an envelope of isoprofit curves for workers with identical going wage rates for all relevant jobs. At a maximum point, the derivative of (3) with respect to the jth job characteristic is proportional to the shadow price of job characteristic j, the latter being the derivative of wage with respect to that job characteristic at the relevant point along the firm's efficiency frontier. Using these conditions and solving provides:[12]

$$\tau_i^\alpha = \tau(\Omega^\alpha; Z_i; G^\alpha; \xi^\alpha), \quad \text{all } i, \alpha \tag{4}$$

where $\tau_i^\alpha = 1$ if this firm wishes to hire α for job i; 0 otherwise.

[7]Note that in Figure 1, $\tilde{w}_i^\alpha$ measures the "money wage plus the monetary equivalent of psychic wage" for person α with indifference curves $\bar{u}^\alpha$. It is clear that unless by chance E_1E_1 is linear, this aggregate form must vary with taste even within a group of workers facing identical wage offers and presumably of similar skills.

[8]Stuart Altman and Robert Barro consider a special case within this class of functions. Note that (2) appeals to McFadden's axiom on the irrelevance of the alternative set effect in the context of nonreplicated experiments to avoid the writing of all job characteristics of all jobs in (2). See McFadden, p. 110, axiom 3.

[9]Thus, high levels of strength endowed in a person probably do not add, and may even detract, from productivity in a purely sedentary job. This is the basis of comparative advantage in job performance. Note that it is not simple to distinguish between "ability" and "working condition" job characteristics: e.g., extreme temperatures in the work place are uncomfortable, but there is in some sense an ability to withstand such extremes.

[10]Since supply of labor conventionally refers to turning up at work, supply of effort on the job is subsumed into productivity on the demand side of the market.

[11]The marginal contribution of any individual to a firm's revenue product is also likely to depend upon the occupational activities of all other workers both because of diminishing marginal productivity, and because the necessity of performing team work may render this worker's productivity dependent upon the qualities of fellow workers in his own and related jobs. This activity vector is omitted from (3) here for it adds nothing to the problem given the assumptions which follow.

[12]Note again our reliance on the irrelevance of the alternative set effect. Otherwise, all wage offers and personal characteristics of every person, as well as all jobs' characteristics, enter (4).

C. *Market Clearing*

A complete equilibrium may be defined by the matching of plans on behalf of employees and employers so that

$$\tau_i^\alpha = \theta_i^\alpha \tag{5}$$

In this equilibrium, the two efficiency frontiers for any worker, such as E_1E_1 and e_1e_1 in Figure 2, must be coincidental in the neighborhood of choice, generating the "kissing" of corresponding isoprofit ($\bar{\pi}^\alpha$) and indifference ($\bar{u}^\alpha$) curves familiar from Rosen (1974). The system of equations (2), (4), and (5) may presumably then be solved for the reduced form equilibrium wage (6) and occupational allocation (7) equations:

$$w_i^\alpha = w(Z_i, G^\alpha, \epsilon^\alpha, \xi^\alpha) \tag{6}$$

$$t_i^\alpha = t(Z_i, G^\alpha, \epsilon^\alpha, \xi^\alpha) \tag{7}$$

To the set of equations (6), the term "hedonic wage equations" may be applied.[13]

FIGURE 2

II. Hedonic Wage Equations Estimated

A. *Data Sources*

In order to estimate the reduced form wage equation in (6) it is desirable to have observations of individual's wage rates, personal characteristics, and job characteristics. Such a data file is compiled for this study by combining information from two separate *U.S.* sources:

a) The *Survey of Economic Opportunity* (*SEO*) reports hourly earnings and various personal data for each adult in the survey. The *SEO* actually comprises two separate half samples: a national random sample; and a supplementary sample, drawing from predominantly nonwhite neighborhoods. Observations are confined in this study to whites from the random half sample and blacks from the other half sample, there being too few blacks in the former.

b) A *Dictionary of Occupational Titles* (*DOT*) tape provides the data on job characteristics, describing in terms of attributes the task involved in each of 13,778 occupations from the *DOT* classification.

To generate a suitable data file, the two separate sources must be spliced. Unfortunately, the *SEO* in reporting adults' occupations does not adopt the *DOT* classification, but rather the *U.S.* 1960, five-digit, census classification scheme (with 295 categories). A weighted cross-classification table of the two occupational coding schemes, prepared by the U.S. Department of Labor, enables one to link the separate *SEO* and *DOT* files. The cross-classification table is a matrix, each element being the number of adults from the October 1966 *Current Population Survey* (35,000 households) who enter a particular job cross-classification cell. From the resulting 13,778 by 295 matrix it is possible to compute the sample probabilities of performing each of the *DOT* occupations given census occupation.[14]

[13] The equilibrium condition (5) may rightly be considered too exacting in its informational content. An alternative view is to imagine employers predicting π^α with (3), using observed values of $\hat{G}^\alpha$ (see Michael Spence, 1973). Under risk neutrality, and assuming ϵ^α and ξ^α to be additive in (1) and (3) with zero means, equilibrium may be defined by $E(\tau_i^\alpha) = E(\theta_i^\alpha)$ for workers with common $\hat{G}^\alpha$. This condition provides a reduced form similar to (6) and (7), but in terms of $\hat{G}^\alpha$, with ϵ^α, ξ^α omitted.

[14] Two of the census occupations (laborers n.e.c. and operatives and kindred workers n.e.c.) contain a sub-

Job attributes on the *DOT* are measured as dummy or step value variables. Converting the latter to dummies, one can compute the probability of holding a *DOT* job with an attribute at a particular level given census occupation (industry). Using adults' reported census occupation (industry), a vector of job attribute probabilities is then assigned to each adult in the *Survey of Economic Opportunity*.[15] Unfortunately, 17 percent of the *SEO* working adults failed to report their occupation for the week of the survey.[16] The missing job characteristics for this group are therefore predicted here by use of equations regressing job on personal characteristics for the remaining group.[17]

B. *Specification and Estimation*

The following variables are those included in the estimated relation:

$\hat{w}_i^\alpha$ = hourly earnings deflated by a locational cost-of-living index[18]

$\hat{g}_i^\alpha$ = age in years

$\hat{\mu}^\alpha$ = 1 if adult is a union member; 0 otherwise

$\hat{z}_{ij}$ = probability of holding a *DOT* job with the following characteristics:

j = 1. specific vocational preparation (*SVP*), required for average job performance, in excess of one month.
2. higher levels of general educational development required (*GED*, an indicator of reasoning development required of a worker for average satisfactory task execution).[19]
3. supervising a group of workers.
4. nonsedentary (i.e., task does not involve both mostly sitting and never lifting more than ten pounds.)
5. repetitive or short-cycle operations carried out according to set procedures or sequences.
6. a work environment embracing at least one of the physical conditions: extremes of heat or cold; wet or humid conditions; sufficient noise to cause marked distraction or possible injury to the sense of hearing; definite risk of bodily injury; fumes, odors, toxic conditions, dust, or poor ventilation.[20]

The specification chosen is:

$$ln\,\hat{w}_i^\alpha = \beta + \phi(\hat{g}_i^\alpha) + \beta_0\hat{\mu}^\alpha + \sum_{j=1}^{6} \beta_j \hat{z}_{ij} + e^\alpha \tag{8}$$

where ϕ is a piece-wise linear function with break-points chosen at ages 14, 25, 55, and 99, the third being omitted; and e^α is a stochastic disturbance term.

Note that the $\hat{z}_{ij}$ measure true job attributes with error, but, being the census cell means, it is reasonable to suppose an absence of bias, though some degree of heteroskedasticity is incurred in using ordinary least squares.[21] A separate equation (8) is estimated for the four race/sex groups for

stantial proportion of the labor force and are particularly heterogeneous with respect to type of work done. These two categories are therefore subdivided here by five-digit census industry, thereby reducing the variance of job characteristics within each cell. The *DOT* job probabilities within industry subcells are then computed by matching the industry associated with each job on the *DOT* tape and the very detailed census industrial classification.

[15]One outcome of this process is that since only one cross-classification matrix exists for all race/sex groups, it has to be assumed that the distribution over *DOT* occupations within a census cell is the same for each race/sex group.

[16]The *SEO* in fact asks occupation for most of 1966 rather than the week of the survey. The 17 percent figure therefore includes persons not responding positively to having retained the same job, as well as those failing to report 1966 occupation (industry).

[17]These equations are tabulated in full in the author (1972), Appendix A, and summarized in the author (1974).

[18]These indices are condensed from Bureau of Labor Statistics (1967). Separate indices are used for the 12 large *SMSA*s, and for other *SMSA* and non-*SMSA* locations within the four major regions. See the author (1972), pp. 204–05.

[19]"Higher" refers to *GED* levels 4 and above. Richard Eckaus tentatively suggests that level 4 might be associated with high school completion (after converting Eckaus' 7 classification to the 6 in this later edition of the *DOT*).

[20]For more detailed information on all job characteristics, see *DOT*, Appendix.

[21]Given a large number of cells, the variance of errors due to errors-in-variables is probably small compared to other sources, and the loss of efficiency from not employing weighted least squares therefore small.

TABLE 1—ESTIMATED HEDONIC WAGE EQUATIONS

	White Males			Black Males			White Females			Black Females		
Schooling	0–8	9–11	12	0–8	9–11	12	0–8	9–11	12	0–8	9–11	12
Constant	.470	.765	.819	−.111	.427	.284	.350	.603	.413	.353	.245	.287
	(.106)	(.086)	(.076)	(.176)	(.149)	(.115)	(.160)	(.119)	(.070)	(.278)	(.129)	(.095)
Age 14	−.570	−.819	−.690	−.469	−.573	−.390	−.183	−.380	−.310	−.380	−.237	−.343
	(.057)	(.042)	(.050)	(.067)	(.054)	(.079)	(.097)	(.070)	(.047)	(.123)	(.063)	(.072)
25	−.101	−.136	−.127	−.123	−.186	−.064	−.064	−.048	−.037	−.032	−.077	−.168
	(.038)	(.035)	(.026)	(.042)	(.048)	(.052)	(.076)	(.073)	(.035)	(.061)	(.054)	(.053)
99	−.687	−1.162	−.772	−.754	−1.143	−1.474	−.979	−1.371	−.423	−.208	−.257	−.700
	(.118)	(.191)	(.170)	(.162)	(.411)	(.573)	(.209)	(.359)	(.216)	(.180)	(.357)	(.475)
Union member	.299	.239	.160	.447	.268	.253	.262	.218	.208	.324	.235	.125
	(.022)	(.021)	(.017)	(.027)	(.029)	(.030)	(.051)	(.055)	(.032)	(.060)	(.050)	(.042)
SVP	.344	.228	.335	.336	.197	.335	.246	.020	.166	.201	.322	.433
	(.059)	(.057)	(.065)	(.056)	(.062)	(.062)	(.071)	(.077)	(.053)	(.071)	(.056)	(.052)
GED	.423	.245	.241	.444	.325	.139	.382	.285	.315	.128	.150	.319
	(.050)	(.046)	(.039)	(.067)	(.070)	(.068)	(.132)	(.103)	(.047)	(.148)	(.097)	(.075)
Supervise	.151	.227	.152	.184	.159	−.128	−.440	−.185	−.067	.019	−.348	.045
	(.086)	(.068)	(.053)	(.162)	(.154)	(.208)	(.342)	(.245)	(.194)	(.344)	(.257)	(.201)
Nonsedentary	−.260	−.138	−.170	.055	−.055	.275	−.504	−.398	−.188	−.610	−.217	−.282
	(.092)	(.068)	(.038)	(.172)	(.135)	(.080)	(.139)	(.089)	(.031)	(.272)	(.111)	(.064)
Repetitive	.296	.119	.103	.450	.264	.077	.400	.274	.223	.152	.110	.256
	(.045)	(.045)	(.043)	(.046)	(.056)	(.059)	(.082)	(.080)	(.049)	(.069)	(.065)	(.071)
Physical	.133	.100	.068	.109	.128	−.077	.149	.198	.033	.377	.121	.195
conditions	(.039)	(.035)	(.028)	(.047)	(.048)	(.051)	(.075)	(.079)	(.054)	(.070)	(.061)	(.066)
Degrees of freedom	1602	1758	2699	1625	956	757	735	990	2020	1038	809	826
Sum squared residuals	264.9	258.4	392.2	326.6	143.1	101.1	203.9	337.9	351.2	246.9	134.6	125.4
Residual variance	.165	.147	.145	.201	.150	.134	.278	.341	.174	.238	.166	.152
R^2	.374	.461	.199	.330	.319	.165	.233	.149	.144	.146	.152	.230

each of three levels of schooling: 0–8, 9–11, and 12 grades of school completed.[22] Persons failing to report a wage are omitted, as are those reporting no wage, the latter because leisure activities and their attributes are not known. Also, those adults from families with family business or farm income exceeding $1000 are excluded, there being considerable difficulty in distinguishing labor from property incomes in such cases.

C. *Results*

The estimated coefficients, with their standard errors in parentheses beneath, are given in Table 1.[23] The coefficients on the constant term are estimates of mean *log* wage for 55 year old nonunion workers possessing jobs with none of the included job characteristics. Of the race/sex classes within this group, only white males demonstrate monotonic increases in wage rate as grades of schooling completed rise. Note that being black and being female serve to reduce the constant term's coefficient, indicating that this well-known phenomenon is not a consequence of the included job attributes. Similarly, the familiar profile of wages across age groups is observed—at first rising, then turning down—though for females there is no significant

[22]Regressions for higher education categories are not reported here owing to space limitations and because very few observations on blacks at these levels are available.

[23]Reestimating these equations, omitting observations with missing job characteristic values predicted by regression equations, generates no significant changes.

rise from ages 25 to 55. It is interesting that the profiles persist here despite the inclusion of a measure of vocational preparation required for the job, a result which proves insensitive to alternative measures of the specific vocational preparation variable.

In my 1974 paper I report that, *ceteris paribus*, union membership is associated with those job characteristics more common to lower levels of schooling. However, it is seen from Table 1 that union membership has a very substantial effect on wages, even holding job attributes steady, this partial effect reaching a level of 45 percent for black males with low levels of schooling. Within each sex, the effect of union membership is towards equality of the races, though not such as to entirely offset constant term discrepancies. In contrast, union membership serves to widen sex differentials for given job types within each race.

James Scoville fails to discover any "significant" relationship between occupational wages and specific vocational preparation, whereas the corresponding coefficients in Table 1 are quite large. They indicate a wage differential in excess of 25 percent between groups with more and less than one month of training (with the notable exception of female whites). Similarly, within each class (except black females with lower schooling levels), those persons paired with jobs demanding higher levels of general educational development receive significantly greater wages, even though number of grades complete is held constant. (These upper steps of *GED* are, incidentally, highly correlated with the *Dictionary of Occupational Titles* measures of intelligence required to perform a job.)

White males' wages improve some 15 to 20 percent upon assignment to a supervisory job, whereas women of both races tend, if anything, to lose. The small number of black males and women who supervise is reflected in their high standard errors on this coefficient. The nonsedentary jobs are those requiring more lifting and physical exertion. The undertaking of such tasks, if distasteful, is not rewarded, and neither is the possession of any physical ability which might be associated with heavier jobs, but rather the converse, which suggests the omission of some skill associated with sedentary job holders.

Indeed, each of these first four job characteristics is probably associated with shifts of the efficiency frontiers rather than movements along them. However, the positive coefficients on the repetitive and physical conditions attributes might well be viewed as compensatory payments, offering some of the first evidence in support of Adam Smith's notion of equalizing differences, and reversing the negative simple correlation between these characteristics and wage rates.

III. Psychic Wages in the Returns to Schooling

A. *Theory*

Suppose the wage equation (6) may be rearranged as:

$$(9)\quad ln(w_i^a + \Psi^a(Z_i)) = P'G^a + \epsilon^a + \xi^a$$

where P is a vector of scalars $\{p_m\}$. This, of course, is the standard human capital wage equation, for in equilibrium all elements of Z_i in (6) are arguments of the utility function so $\Psi^a(Z_i)$ may be termed the monetary equivalent of psychic wage. However, when estimating the rates of return to malleable elements of G^a, it is common practice to neglect Ψ^a in measuring the dependent variable. What is the consequence of this omission?

Writing the commonly estimated rate of return to investment in element g_m^a of G^a as $\hat{p}_m$, which equals $\partial(log\ w_i^a)/\partial g_m^a$, it is easily seen that:

$$(10)\quad \hat{p}_m - p_m = \frac{\partial\, ln[w_i^a/(w_i^a + \Psi^a(Z_i))]}{\partial g_m^a}$$

In words, the direction of error from omission of psychic wages depends upon whether the share of money wages in total compensation increases or decreases with the relevant personal characteristic.

The money wage plus the monetary equivalent of psychic wage measured in (10) is, of course, an index number, using

TABLE 2—PSYCHIC WAGES IN THE RETURNS TO SCHOOLING

Schooling		$\hat{\rho}_m - \rho_m$			$(\hat{\rho}_m - \rho_m)/\hat{\rho}_m$			ρ_m		
		Computed with shadow prices from schooling class								
Classes	Sex/Race	0-8	9-11	12	0-8	9-11	12	0-8	9-11	12
9-11, 12	*MW*	-.034	-.018	-.017	-.334	-.180	-.169	.134	.119	.117
	B	-.034	-.025	.000	-.421	-.309	.002	.114	.105	.080
	FW	-.039	-.036	-.023	-.392	-.361	-.233	.139	.136	.123
	B	-.027	-.015	-.043	-.265	-.147	-.423	.129	.117	.145
12, 13-16	*MW*	-.023	-.013	-.012	-.301	-.168	-.158	.100	.090	.089
	B	-.048	-.035	-.002	-.522	-.378	-.019	.140	.127	.094
	FW	-.035	-.029	-.024	-.382	-.313	-.261	.127	.121	.116
	B	-.035	-.020	-.055	-.240	-.138	-.372	.182	.167	.202
13-16, 17+	*MW*	-.010	-.005	-.005	-.272	-.146	-.141	.048	.043	.043
	B	-.026	-.019	-.001	-.241	-.178	-.013	.134	.127	.109
	FW	-.017	-.014	-.012	-.157	-.130	-.110	.126	.123	.121
	B	-.012	-.007	-.019	-.082	-.052	-.136	.153	.149	.161

shadow prices to add the "quantities" of money and psychic wages. At any one level of schooling there exist at a point in time many shadow prices that might be used for aggregation, as the efficiency frontier is not generally linear, but since the rate of return to schooling is itself an average, the mean shadow prices are obvious candidates. Even these means vary across levels of schooling, and the distinct efficiency frontiers corresponding to different educations normally acquire different shapes, as E_1E_1 and E_2E_2 in Figure 2. Clearly, just one set of shadow prices must be adopted if a meaningful index is to be constructed, though the index and consequently the computed values in (10) are not independent of this selection. Further, consider the notion of a "true" index of profitability from investment in schooling, defined as the money wage compensation necessary on average to leave the efficiency frontiers for two levels of schooling just touching the same indifference curve for every person. Unfortunately, as John Muellbauer shows, such true indices generally stand in an unknown relationship to the Laspeyres and Paasche counterpart.[24] Thus, distance XY in Figure 2 (resulting from a parellel shift of E_1E_1 to $E_1'E_1'$) may either exceed or be less than the equivalent variation measure of change in $\tilde{\omega}_i^\alpha$ constructed from the shadow prices at X on E_1E_1.

[24]See also N. Anders Klevmarken.

B. *Results*

This last subsection computes the error, $\hat{\rho}_m - \rho_m$, accruing from neglect of changes in the repetitive nature of work and the physical conditions of the work environment, when evaluating the rate of return to schooling. Three alternative calculations are performed for each race/sex group, each using the relevant coefficients on these job characteristics taken from a particular level of schooling in Table 1. Rates of return and the errors are allowed to vary in step form by pairwise comparison of levels of schooling 9-11, 12; 12, 13-16; 13-16, 17+.[25]

The results in Table 2 show that $\hat{\rho}_m - \rho_m$ is negative at all levels of schooling, independent of the shadow prices chosen. This evidence suggests that omission of psychic wages does result in underestimation of the average rate of return to schooling. It occurs because wages represent a falling frac-

[25]The $\hat{\rho}_m$ is here computed very simply as the difference in mean *log* wage rate divided by difference in mean grades complete between these pairs:

$$\hat{\rho}_m = E\left[\frac{ln\ \hat{w}_i^\alpha - ln\ \hat{w}_i^{\alpha'}}{g_m^\alpha - \hat{g}_m^{\alpha'}}\right]$$

where α and α' are members of adjacent schooling levels. The ρ_m is calculated in a similar fashion, measuring $\hat{w}_i^\alpha$ by:

$$w_i^\alpha + \Psi^\alpha(Z_i) = \hat{w}_i^\alpha\left[1 - \sum_{j=1}^{6}\hat{\beta}_j\hat{z}_{ij}\right]$$

tion of total compensation as schooling rises. It is not possible, however, to state this conclusion definitely, for the consequences of including other components of psychic wages remain uncertain.

The computed errors in Table 2 tend to be larger as shadow prices from lower schooling levels are employed, but how big are these errors? The second vertical panel in Table 2 shows $(\hat{\rho}_m - \rho_m)/\hat{\rho}_m$. Almost nowhere is the error less than 15 percent of the commonly estimated rate of return, and actually exceeds one third in a number of cases. Typically, the errors are smallest for the returns to graduate school. Since no clear race/sex patterns seem to emerge in these errors, the estimates of ρ_m reported in the last panel tend to preserve the rank orderings of the $\hat{\rho}_m$.

IV. Closing Remarks

This paper demonstrates the feasibility of a multiple attribute approach to the problem of psychic wages and provides some of the first systematic empirical support for Adam Smith's notion of equalizing wage differentials. Thus, it is shown that, *ceteris paribus*, workers do receive substantially higher money wages in compensation for undertaking jobs embracing repetitive routines and obnoxious physical work environments. Further, it is established that tasks associated with higher levels of specific vocational preparation and general educational development, and (for white males) supervisory jobs, do pay considerably higher wages *ceteris paribus*, suggesting reward to some omitted set of skills not fully reflected in the common schooling and age variables.

The results presented here on psychic wages in the return to schooling indicate a considerable downward bias from estimating such returns in terms of monetary rewards alone. In essence, this result follows from the inference that the pecuniary fraction of total compensation is a declining function of schooling for all race/sex groups.

Clearly, it would be of interest to proceed in future work to the estimation of the supply (2) and demand (4) side forces which lead to the reduced form compensations studied above. However, the identification problems inherent in such an exercise are far from trivial, for, as Section I indicates, both job and personal characteristics generally play a role on both sides of the market.

Appendix: The Notation System

Subscript i indicates a value associated with job i. Superscript α indicates a value associated with person α. A bar indicates a constant value. A "hat" indicates a measured counterpart. A tilde indicates inclusion of monetary equivalent of psychic wages.

$Z_i = [z_{ij}]$ are job characteristics.

$G^\alpha = [g_m^\alpha]$ are personal characteristics.

$\Omega^\alpha = [\omega_i^\alpha]$ are wage offers.

$P = [\rho_m]$ are returns to personal characteristics.

$\tau_i^\alpha = 1$ if α is selected for job i, 0 otherwise.

$\theta_i^\alpha = 1$ if α selects job i, 0 otherwise.

$t_i^\alpha = 1$ if α is selected for and selects job i, 0 otherwise.

$\epsilon^\alpha, \xi^\alpha, e^\alpha =$ idosyncratic, unobserved elements in the utility, net profit, and wage regression functions, respectively.

$w_i^\alpha =$ market-clearing value of ω_i^α.

$\mu^\alpha =$ union membership dummy.

$u^\alpha =$ utility.

$\pi^\alpha =$ net profit.

REFERENCES

S. H. Altman and R. J. Barro, "Officer Supply—The Impact of Pay, the Draft, and the Vietnam War," *Amer. Econ. Rev.*, Sept. 1971, *61*, 649–64.

M. J. Boskin, "A Conditional Logit Model of Occupational Choice," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, 389–98.

R. S. Eckaus, "Economic Criteria for Education and Training," *Rev. Econ. Statist.*,

May 1964, *46*, 181–90.
Z. Griliches, "Notes on the Role of Education in Production Functions," in W. Lee Hansen, ed., *Education, Income and Human Capital*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 35, New York 1970.
N. A. Klevmarken, "A Note on New Goods and Quality Changes in the True Cost-of-Living Index in View of Lancaster's Model of Consumer Behavior," *Econometrica*, Jan. 1977, *45*, 163–73.
K. J. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–57.
R. E. B. Lucas, "Working Conditions, Wage-Rates and Human Capital: A Hedonic Study," unpublished doctoral dissertation, M.I.T., Oct. 1972.
———, "The Distribution of Job Characteristics," *Rev. Econ. Statist.*, Nov. 1974, *56*, 530–40.
———, "Hedonic Price Functions," *Econ. Inquiry*, June 1975, *13*, 157–78.
D. McFadden, "Conditional Logit Analysis of Qualitative Choice Behavior," in Paul Zarembka, ed., *Frontiers in Econometrics*, New York 1974.
J. Muellbauer, "Household Production Theory, Quality, and the 'Hedonic Technique'," *Amer. Econ. Rev.*, Dec. 1974, *64*, 977–94.
S. Rosen, "On the Interindustry Wage and Hours Structure," *J. Polit. Econ.*, Mar./Apr. 1969, *77*, 249–73.
———, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 34–55.
James G. Scoville, *The Job Content of the U.S. Economy 1940–1970*, New York 1970.
A. M. Spence, "Job Market Signaling," *Quart. J. Econ.*, Aug. 1973, *87*, 355–74.
J. Tinbergen, "On the Theory of Income Distribution," *Weltwirtschaftliches Archiv*, 1956, *77*, 155–73.
F. R. Welch, "Linear Synthesis of Skill Distribution," *J. Hum. Resources*, Summer 1969, *4*, 311–27.
U.S. Bureau of Labor Statistics, *Three Standards of Living*, Washington, Spring 1967.
U.S. Department of Labor, Manpower Administration, *Dictionary of Occupational Titles, (DOT)* Vol. 2, Washington 1965.
U.S. Office of Economic Opportunity, "*Survey of Economic Opportunity*," (*SEO*) conducted spring 1967, available on tape, Data Bank, Univ. Wisconsin.

Homothetic and Non-Homothetic *CES* Production Functions

By RYUZO SATO*

In economic theory the production function is generally a concept stating quantitatively the purely technological relationship between the output and the inputs of factors of production. An essential purpose of the concept is to describe the substitution possibilities among the factors of production in order to achieve a given level of output. In the past it has been convenient to approximate the relationship by a special class of functions containing such specific forms as the Cobb-Douglas and the (homothetic) *CES* (constant elasticity of substitution). It is no exaggeration to say that these specific forms are the most frequently used nonlinear type of special functions in the field of economic analysis. They have been construed as the simplest type of meaningful functions in both theoretical and empirical studies of production relationships (see Kenneth Arrow et al., Daniel McFadden, and Jacob Paroush) and have received privileged attention in economic growth analysis (see the author, 1970).

In my previous papers (1974, 1975a) it is shown, however, that there exists a more general and more meaningful class of *CES* production functions, i.e., non-homothetic *CES* functions, which include the ordinary (or homothetic) *CES* or the Cobb-Douglas functions as special cases. While these earlier papers (especially 1975a) focused on the mathematical aspects of the non-homothetic *CES* functions, I shall, in this paper, endeavor to present some economic interpretations of the properties of the functions together with some results of empirical applications.

I shall first provide some economic justifications as to why such non-homothetic *CES* functions may be useful in production analysis.[1] Consider a typical estimation problem of production functions under the competitive markets and the homogeneity or homotheticity assumptions. If the underlying production function is of the *CES* type together with the assumption of Hicksian neutral technical progress, then the marginal rate of substitution between the two factors, capital and labor, ω, and the capital-labor ratio, $k = K/L$, are related according to the equation,

$$(1) \quad \log k = \log a + \sigma \log \omega, \quad \sigma = \text{constant}$$

where σ is the elasticity of substitution. Because of the homotheticity assumption, the capital-labor ratio is *independent* of the level of output and of the neutral type of technical progress. It simply depends on the marginal rate of substitution, or alternatively on the relative factor prices. Empirical data (time-series in particular), however, suggest that the factor ratio varies even at a constant price ratio.

The well-known technique to deal with this kind of situation is the introduction of biased technical progress, specifically the so-called factor-augmenting technical progress (for example see the author, 1970). But this device is often of no use due to the impossibility (theorem) of identification of the bias and of the substitution effect. This problem can be resolved if we relax the homogeneity or homotheticity assumption (hereafter referred to only as "homotheticity" assumption), so that *the level of output and the degree of technical progress will explicitly have effects on factor combinations.* The empirically convenient form for the marginal rate of substitution/capital-labor ratio relationship (*MRS*-K/L relationship) is thus:

*Professor of economics, Brown University. I wish to acknowledge financial assistance from the National Science Foundation and the Guggenheim Foundation. For helpful comments, thanks are due to the managing editor and an anonymous referee.

[1]Justifications for demand analysis are given in the papers by Paul Samuelson and the author (1976).

$$(2) \quad \log k = \log a + \sigma \log \omega + b \log Y + c \log T(t)$$

where Y is the level of output and $T(t)$, the index of technical progress. The underlying production function which is consistent with the above relationship is, in fact, the non-homothetic class of *CES* functions, the properties of which we propose to interpret.

I. Properties of the Non-Homothetic Family of *CES* Functions

I shall begin by first presenting the general expression for the *non-homothetic family of CES functions*[2] (hereafter referred to as *NH-CES* as opposed to *H-CES*, homothetic *CES*):

$$(3) \quad \text{(i)} \quad F(K, L, Y) = C_1(Y)K^{-\rho} + C_2(Y)L^{-\rho} - 1 = 0, \quad \text{for} \quad \sigma \neq 1$$

$$\text{(ii)} \quad F(K, L, Y) = C_1(Y)\log K + C_2(Y)\log L - 1 = 0, \quad \text{for} \quad \sigma = 1$$

where $\rho = (1 - \sigma)/\sigma$ and σ = the elasticity of substitution. Sometimes it is convenient to write (3) as

$$(3') \quad \text{(i)} \quad F(K, L, Y) = K^{-\rho} + C(Y)L^{-\rho} - H(Y) = 0, \quad \sigma \neq 1$$

$$\text{(ii)} \quad F(K, L, Y) = \log K + C(Y)\log L - H(Y) = 0, \quad \sigma = 1$$

Rather than presenting two expressions, one for $\sigma \neq 1$ and one for $\sigma = 1$, it is much neater and more compact to write the general family of *NH-CES* production functions as

$$(3'') \qquad X_1 + C(Y)X_2 = H(Y)$$

where $X_1 = K^{-\rho}$ and $X_2 = L^{-\rho}$ for $\sigma \neq 1$ or $X_1 = \log K$ and $X_2 = \log L$ for $\sigma = 1$, and both $C(Y)$ and $H(Y)$ are monotone functions of Y.

[2]For the derivation, readers are referred to the articles by the author (1974, 1975a), especially equation (4) or (4a) of the 1975a article.

A. *NH-CES and NH-Cobb-Douglas*

Although equation (3-i) contains all of the *CES* functions, including the Cobb-Douglas function as limiting cases, it would perhaps facilitate understanding if this new family of *CES* functions were compared with the ordinary (or homothetic) *CES* and Cobb-Douglas production functions. Thus, equation (3-i) is the non-homothetic family of *CES* functions, as contrasted with the ordinary *CES* functions; while equation (3-ii) is the non-homothetic counterpart of the Cobb-Douglas family. Let us refer to (3-i) as *NH-CES* and to (3-ii) as the *Non-Homothetic Cobb-Douglas* or *NH-CD* for short.

B. *Properties of NH-CES and NH-CD*

There are a number of specific properties that are unique to the non-homothetic production functions:

1. The Marginal Rate of Substitution and the Non-Homotheticity Parameter

The most distinctive property of *NH-CES* and *NH-CD* is, of course, that the production function is non-homothetic and is characterized by *variable* marginal rate of substitution, even at a *constant* factor ratio. That is to say, unlike the cases of the *H-CES* and the *CD* functions, the expansion path of the isoquant map of *NH-CES* and *NH-CD* production functions is *not* a straight line, but *varies* depending upon the *level* of output. This is due to the fact that the marginal rate of substitution now depends not only on the factor ratio k, but also on the *output level* Y for any given value of the elasticity of factor substitution σ, i.e.,

$$(4) \quad \omega = \frac{\dfrac{\partial Y}{\partial L}}{\dfrac{\partial Y}{\partial K}} = \left(\frac{K}{L}\right)^{1+\rho} C(Y) = k^{1/\sigma} C(Y)$$

where $\partial Y / \partial L$ = marginal product of L

$$= \frac{\rho L^{-(1+\rho)} C(Y)}{C'(Y)L^{-\rho} - H'(Y)} > 0$$

and $\partial Y / \partial K$ = marginal product of K

$$= \frac{\rho K^{-(1+\rho)}}{C'(Y)L^{-\rho} - H'(Y)} > 0$$

The typical isoquant map of a *NH-CES* or *NH-CD* production function is depicted in Figure 1. The expansion path *AB* in Figure 1 is not a straight line, but bends backward or forward depending upon the elasticity of factor substitution. Also, unlike the *H-CES*, the *NH-CES* (also *NH-CD*) functions have lower and upper limits for the effective values of *K* and *L*, as is evident in Figure 1.

The non-homothetic aspect of the production function may be best characterized by the existence of the non-homotheticity coefficient (or parameter) for the marginal rate of substitution. The derivative of $C(Y)$ in the production function (3″) and also in the marginal rate of substitution function, with respect to Y, i.e., dC/dY, may be called the non-homotheticity coefficient. If the behavior of the coefficient may be approximated by a parameter, the unique properties of the isoquant map and of the marginal rate of substitution is summarized by the *non-homotheticity parameter*. Thus, obviously, in the cases of the *NH-CES* and the *NH-CD* production functions, the parameter is *not* zero, i.e., $dC/dY \neq 0$, while in the ordinary *CES* and *CD* functions, it is identically equal to zero, $dC/dY = 0$. Hence, the general class of *CES* production functions contain a *non-homotheticity parameter* in addition to the usual *distribution, substitution, and efficiency parameters*. The ordinary *CES* family is a special case of this general class when the non-homotheticity parameter *ceases* to exist.

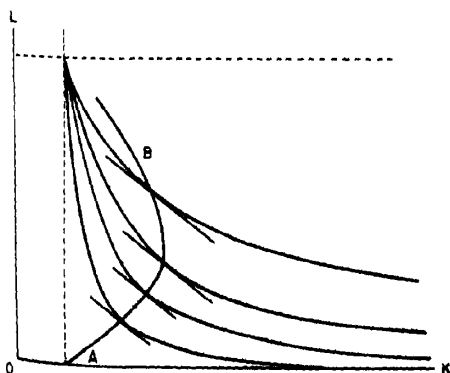



FIGURE 1. ISOQUANT MAP OF *NH-CES* OR *NH-CD*

The non-homotheticity assumption presents no additional problem regarding the convexity of isoquants. The marginal rate of substitution will be diminishing as long as $\rho > -1$ for any positive marginal rate of substitution, as we have $\partial\omega/\partial(L/K) = -(1+\rho)\omega(K/L) < 0$ for $\infty > \rho > -1$ and $\omega > 0$.

2. Substitution and Non-Homotheticity Parameters and Income Distributions

The second significant characteristic of the general family of non-homothetic *CES* and *CD* production functions is that, unlike the ordinary *CES* functions, the behavior of *the distribution of factor incomes depends not only on the substitution parameter but also on the non-homotheticity parameter.* It should be noted that in the case of non-homothetic functions, the sum of factor income shares will *not* add up to the total output, or the absorption theorem does *not* hold, because of the very nature of the non-homotheticity or non-homogeneity assumption. Thus, the behavior of *each* factor's income share is not directly related with the factor ratio nor with the substitution elasticity. However, if the movements of income distributions are expressed in the form of an income shares *ratio*, then the factor ratio and the substitution elasticity play important determining roles. For the non-homothetic *CES* case, the ratio of labor's income to capital's income from (4) is equal to

$$(5) \qquad \frac{\beta}{\alpha} = \frac{\dfrac{\partial Y}{\partial L} \cdot L}{\dfrac{\partial Y}{\partial K} \cdot K} = k^{1/\sigma - 1} \cdot C(Y)$$

where α and β are incomes of capital and labor, respectively, under competition.

Differentiating (5) partially with respect to k and Y, one obtains

$$\text{(6a)} \quad \frac{\partial(\beta/\alpha)}{\partial k} = \left(\frac{1}{\sigma} - 1\right) k^{1/\sigma - 2} \cdot C(Y)$$

$$\text{(6b)} \quad \frac{\partial(\beta/\alpha)}{\partial Y} = k^{1/\sigma - 1} \cdot C'(Y)$$

Thus, it may be stated that, *as long as the substitution elasticity is greater (less) than unity, capital's income relative to labor's income rises (falls) when the capital-labor ratio increases and that, as long as the non-homotheticity coefficient is positive (negative), labor's income relative to capital's income rises (falls) whenever output increases.* Hence, in the non-homothetic Cobb-Douglas family, even though the elasticity of substitution is unitary, the relative income distribution *varies* depending upon the non-homotheticity coefficient (parameter). This behavior of relative income shares is in sharp contrast to the behavior under the usual Cobb-Douglas function.

3. Different Types of *NH-CES* and *NH-CD*

Another characteristic of the non-homothetic family of *CES* and *CD* functions is that there are an infinite number of *different* types of *NH-CES* and *NH-CD* functions, while there is the *only one* type of homothetic functions. This is because in (3″) the function $C(Y)$ can be any arbitrary function of Y, with the exception, of course, that $C(Y)$ must be chosen in such a way that Y satisfies the usual properties of a production function. On the other hand, when C is constant, we get the only case of homothetic *CES* (or *CD*) functions: if C is constant, (3″) is simply equal to $H(Y) = X_1 + CX_2$ or $Y = G[aX_1 + bX_2]$, $G' > 0$, which is the homothetic *CES* (or *CD*) family.

As $C(Y)$ and also $H(Y)$ are essentially arbitrary functions on an a priori basis, the non-homothetic family of production functions may in general be expressed only as *implicit* formulations. That is to say, unless it can be determined that $C(Y)$ and $H(Y)$ are related in some particular manner, Y *cannot* be explicitly expressed as a function of K and L. Of course, this presents no insurmountable difficulty from the point of view of both theoretical and empirical production analyses. In Section II we shall study several special types of non-homothetic *CES* and *CD* functions, some of which are quite useful for empirical analysis and yield explicit formulations. Incidentally, the relationship defined by (3″) may always be looked at as an *explicit* formulation of the *capital* (*or labor*) *requirement function*, i.e.,

$$X_1 = H(Y) - C(Y)X_2 = R_1(Y, X_2)$$

or

$$X_2 = \frac{1}{C(Y)}(H(Y) - X_1) = R_2(Y, X_1)$$

It defines *explicitly* the amount of capital (or labor) required to produce a given level of output in cooperation with a given amount of labor (or capital). For any given amount of the other factor input, the requirement function R_i ($i = 1, 2$) must be an increasing function of Y: $\partial R_i/\partial Y > 0$. Furthermore, the marginal requirement $\partial R_i/\partial Y$ must also be increasing due to the law of diminishing marginal productivity, i.e., $\partial^2 R_i/\partial Y^2 > 0$.

4. Classifications

The family of *NH-CES* and *NH-CD* production functions may be classified in a number of different ways depending upon the specific purposes in mind. For instance, it may be classified into the separable vs. nonseparable types (see the author, 1974) or into the explicit vs. implicit types (the author, 1975a). It is well known that the ordinary *CES* (or *CD*) type belongs to the explicit and separable class of *CES* functions.

It has been shown (the author, 1974) that the *separable* type of *NH-CES* (or *CD*) can *always* be written as

$$\text{(7)} \qquad Y = F\left(\frac{\beta_1 X_1 + \theta_1}{\beta_2 X_2 + \theta_2}\right)$$

where X_1 and X_2 are the values defined in (3″). An example of nonseparable but explicit types of *NH-CES* (or *CD*) functions may be given as:

$$(8) \qquad Y = F\left[\frac{-X_1 + \sqrt{X_1^2 + 4X_2}}{2X_2}\right]$$

where X_1 and X_2 are again the values defined in (3″).

Another useful way of classifying the family of *NH-CES* (or *CD*) is to consider the *form* of the marginal rate of substitution function. As equation (4) shows, the *form of ω completely depends on the form of $C(Y)$.* We may call the class of *CES* (or *CD*) functions which have the *same form* of the marginal rate of substitution (*MRS*) function, as the *iso-MRS* family of *CES* (or *CD*) functions. Thus, for instance, if $C(Y) = Y^{\delta}$, the *CES* (or *CD*) production functions which have the same form of the marginal rate of substitution $\omega = k^{1/\sigma} Y^{\delta}$, belong to the iso-*MRS* family of the *constant nonhomotheticity parameter*, i.e.,

$$(9) \qquad X_1 + Y^{\delta}X_2 = H(Y)$$

It should be noted, however, that the same form of the marginal rate of substitution does *not* imply the same form of *CES* (or *CD*) functions, except in the case of homothetic *CES* (or *CD*) functions. The form of the *NH-CES* (or *CD*) functions is determined not only by the marginal rate of substitution, i.e., by $C(Y)$, but also by the form of $H(Y)$. If, for instance, $H(Y) = aY^{\delta} + b$, then Y has the form

$$Y = \left(\frac{X_1 - b}{a - X_2}\right)^{1/\delta}$$

which is the same form as (7), the separable type. On the other hand, if $H(Y) = 1/Y^{\delta}$, then Y has the form

$$Y = \left(\frac{-X_1 \pm \sqrt{X_1^2 + 4X_2}}{2X_2}\right)^{1/\delta}$$

which corresponds to (8), the nonseparable type. Consequently, as the forms of (7) and (8) are different, the examples demonstrate that different types of *NH-CES* may have the same form of the marginal rate of substitution function. For the homothetic *CES* (or *CD*) functions $C(Y)$ is constant and thus, the form of the functions is uniquely determined by $H(Y)$: Both the marginal rate of substitution and the production function have the same form, except, of course, for the degree of homotheticity.

Finally, the most precise way of classifying the family of *NH-CES* (or *CD*) functions is based on the classification of the differential equation of the second-order stating the constancy of the elasticity of substitution. Since it is not the purpose of this article to present a formal and technical analysis, I shall simply summarize the results I obtained elsewhere (1975b,c). The differential equation which defines the constancy of the elasticity of substitution and, therefore, which provides the sole basis for the whole family of *NH-CES* (or *CD*) production functions, may be written as

$$(10) \qquad \sigma KL \frac{d^2L}{dK^2} + \frac{dL}{dK}\left(L - K\frac{dL}{dK}\right) = 0$$

$$\sigma = \text{constant} \qquad +\infty > \sigma > 0$$

It is shown that the above differential equation is invariant under the so-called general "projective group" of transformations and that the family of the functions (*CES* or *CD*) generated from (10) belongs to the class of "(general) projective homothetic" production functions, of which the ordinary homothetic *CES* class is a special case.[3] Thus, the whole family of *NH-CES* (or *CD*) may be classified according to the classification of the projective homothetic functions. As this requires a rather advanced knowledge of the group theory of continuous transformations, we shall not involve ourselves with this technical discussion, but simply take up the special case of the so-called "almost homothetic" class in the next section.

[3]By defining $K^{-\rho} = u$ and $L^{-\rho} = v$ (or $\log K = u$ and $\log L = v$ for $\sigma = 1$), equation (10) can be rewritten as $d^2v/du^2 = 0$. Thus, the differential equation is invariant under the "general projective group" whose infinitesimal transformation is $U = (e_1K + e_2L + e_3 - e_7K^2 - e_8KL)\,\partial/\partial K + (e_4K + e_5L + e_6 - e_7KL - e_8L^2)\,\partial/\partial L$. The interested reader should refer to the author (1975b,c).

II. Special Types of *NH-CES* and *CD* Production Functions

A. *Almost-Homothetic CES*

Let us first define the class of *almost-homogeneous* production functions. *A production function is said to be "almost homogeneous" if output rises by γ percent whenever capital and labor increase by γ_1 and γ_2 percent, respectively,*[4] i.e.,

$$(11) \quad Y = f(\lambda^{\gamma_1}K, \lambda^{\gamma_2}L) = \lambda^{\gamma}f(K,L) \quad \text{for } \lambda > 0$$

It is shown (see the author, 1975b) that the function is almost homogeneous if and only if:

$$(11') \quad Y = K^{\gamma/\gamma_1}Q(L^{1/\gamma_2}/K^{1/\gamma_1}) = L^{\gamma/\gamma_2}P(K^{1/\gamma_1}/L^{1/\gamma_2})$$

Obviously if $\gamma_1 \equiv \gamma_2$, the above definition reduces to the standard homogeneity condition.

The almost homothetic production function is defined as any monotone increasing function of the "almost-homogeneous" production function, i.e.,

$$(12) \quad Y = F[f(\lambda^{\gamma_1}K, \lambda^{\gamma_2}L)] = F[\lambda^{\gamma}f(K,L)], \quad F' > 0$$

The isoquant map of an almost-homogeneous production function is identical with the isoquant map of any production function of the almost-homothetic type, except for the labeling of isoquants. Like the isoquant map of the "homothetic" production function, the isoquant of the "almost-homothetic" production function yields some unique properties. It is well known that the isoquant map of the homothetic production function measured in terms of K and L has *straight-line expansion paths.* It can be shown that the isoquant map of the almost-homothetic production function measured *in terms of log K and log L* rather than K and L has *straight-line expansion paths* (see Figure 2). In Figure 2 the axes are measured in terms of *log K* and *log L* and the slope of an isoquant is then nothing but the ratio of distributive shares of capital and labor. The marginal rate of substitution of any almost-homothetic production function is expressed as

[4] The reader is asked to refer to the author (1975b) for further investigations of this useful type of production function, in particular in its relation to "neutral" technical progress.

FIGURE 2. ISOQUANT MAPS OF AN "ALMOST-HOMOTHETIC" PRODUCTION FUNCTION

$$(13) \quad -\frac{dL}{dK} = \frac{f_K}{f_L} = K^{\gamma_2-\gamma_1/\gamma_1}T(L^{1/\gamma_2}/K^{1/\gamma_1})$$

and the share ratio under competition is

$$(14) \quad -\frac{dL}{dK}\cdot\frac{K}{L} = \frac{f_K K}{f_L L} = \frac{P_K \cdot K}{P_L \cdot L} = \frac{\text{share of } K}{\text{share of } L} = R(e^{(1/\gamma_2)\log L - (1/\gamma_1)\log K})$$

If capital and labor grow exponentially at the rates of γ_1 and γ_2 percent, i.e., $K = K_0e^{\gamma_1 t}$ and $L = L_0e^{\gamma_2 t}$, then the share ratio R becomes constant. For different levels of output the share ratio remains unchanged ($A = A'$ in Figure 2) as long as K and L change exponentially at the rates of γ_1 and γ_2 percent, respectively. Hence, *the iso-share curves of an almost-homothetic production function expressed in terms of the isoquant map of log K and log L are straight lines, their slopes being all equal to γ_2/γ_1.* A special type of the usual homothetic pro-

duction function has the isoshare curves whose slopes are all equal to unity, $\gamma_2/\gamma_1 \equiv 1$.

Having made these observations, we can now derive the almost-homogeneous and the more general almost-homothetic class of *CES* production functions. The almost-homogeneous type of *CES* production functions has the form

$$(15) \qquad \beta_1 Y^{\gamma_1} K^{-\rho} + \beta_2 Y^{\gamma_2} L^{-\rho} = 1$$

or

$$(15') \quad K^{-\rho} + \frac{\beta_2}{\beta_1} Y^{\gamma_2 - \gamma_1} \cdot L^{-\rho} = \frac{1}{\beta_1} Y^{\gamma_1}$$

It is easily shown that the production function defined by (15) satisfies the condition of almost-homogeneity, for we have,

$$\beta_1(\lambda^\rho Y)^{\gamma_1}(\lambda^{\gamma_1} K)^{-\rho} + \beta_2(\lambda^\rho Y)^{\gamma_2}(\lambda^{\gamma_2} L)^{-\rho}$$
$$= \beta_1 Y^{\gamma_1} K^{-\rho} + \beta_2 Y^{\gamma_2} L^{-\rho} = 1 = \text{constant}$$
$$\text{or} \quad Y = f[(\lambda^{\gamma_1} K)^{-\rho}, (\lambda^{\gamma_2} L)^{-\rho}]$$
$$= \lambda^\rho f(K^{-\rho}, L^{-\rho})$$

Thus, when capital and labor increase at the percentage rates of γ_1 and γ_2, respectively, output will increase at the rate of ρ percent.

The "almost"-homothetic family of *CES* functions is any monotone transformation of (15). Thus, in general, we must have

$$(16) \quad \beta_1[G(Y)]^{\gamma_1} K^{-\rho} + \beta_2[G(Y)]^{\gamma_2} L^{-\rho} = 1$$

where $G(Y)$ is a monotone function of Y. Solving for $G(Y)$ from (16), we have

$$G(Y) = f[\beta_1 K^{-\rho}, \beta_2 L^{-\rho}, \gamma_1, \gamma_2]$$

or

$$Y = F[f(\beta_1 K^{-\rho}, \beta_2 L^{-\rho}, \gamma_1, \gamma_2)], \quad F = G^{-1}$$

The above must satisfy the condition of almost-homotheticity defined by (12).

As stated in the previous section, the almost-homothetic family of *CES* functions is a typical example of the general "projective" homothetic class of functions. Equation (16) is the invariant family of *CES* functions, in which every transformation of the "magnification group" transforms each curve into some curve of the family. Thus, if the technical progress function is factor-augmenting, i.e., $\bar{K} = e^{a_1 t}K$, $\bar{L} = e^{a_2 t}L$, then there always exists one curve which transforms the production function into some other function with increasing returns to scale,[5] i.e.,

$$f(\beta_1 \bar{K}^{-\rho}, \beta_2 \bar{L}^{-\rho}, \gamma_1, \gamma_2)$$
$$= F[f(\beta_1 K^{-\rho}, \beta_2 L^{-\rho}, \gamma_1, \gamma_2)], F' > 0$$

B. *Separable Family of CES*

Another case of special interest is the separable family of *CES* production functions. In my earlier paper (1974), this type is shown to be generally expressed as

$$(17a) \quad Y = F\left[\left(\frac{\beta_1 K^{-\rho} + \theta_1}{\beta_2 L^{-\rho} + \theta_2}\right)^{-c\alpha/\rho}\right] \quad \text{for } \sigma \neq 1$$

$$(17b) \quad Y = F\left[\left(\frac{\beta_1 \log K + \delta_1}{\beta_2 \log L + \delta_2}\right)^{c\alpha}\right] \quad \text{for } \sigma = 1$$

$$\theta_i = (\alpha - \beta_i)\rho - \beta_i$$
$$\text{and} \quad \delta_i = \beta_i - \alpha \quad (i = 1, 2)$$

The parameters in the above formulation may be called:

$\beta_1 = (1 + \beta_2) = \beta$ = the distribution parameter

$1/\alpha = \gamma$ = the non-homotheticity parameter

$\rho = (1 - \sigma)/\sigma$ = the substitution parameter.

It can easily be shown that when the non-homotheticity parameter ceases to exist ($\gamma \to 0$), (17a) will reduce to the ordinary *CES* type expressed by $Y = F[\beta_1 K^{-\rho} + \beta_2 L^{-\rho}]$, and that when the substitution parameter approaches zero, (17a) will reduce to the class of *NH*-Cobb-Douglas functions defined by (17b).

Furthermore, in equation (17b), when the non-homotheticity parameter converges to zero, the production function will approach to the homothetic family of the ordinary Cobb-Douglas functions.

[5]See the author (1975b) for a detailed discussion of group transformation.

C. *Iso-MRS CES Family of the Constant Non-Homotheticity Parameter*

Finally we present the most general type of the special classes of *CES* functions, which will, therefore, include all the special classes thus far discussed. The classes of almost-homogeneous *CES* and separable *CES* functions will belong to

$$K^{-\rho} + \beta Y^{\delta} L^{-\rho} = H(Y) \tag{18}$$

where δ = the non-homotheticity (constant) parameter. Obviously, when $H(Y)$ takes the form $(1/\beta_1)Y^{-\gamma_1}$, equation (18) will reduce to the almost-homogeneous family, and when $H(Y)$ is equal to $aY^{\delta} + b$, then (18) will reduce to the separable family, for in this case we will have

$$Y^{\delta} = \frac{b - K^{-\rho}}{\beta L^{-\rho} - a}$$

One advantage of the above formulation is that the marginal rate of substitution may be expressed in a form for which ordinary econometric techniques can be utilized to estimate the parameters of the *CES* family. As a matter of fact, equation (2) at the beginning of this paper, neglecting the technical progress term, is nothing but the equation resulting from the marginal rate of substitution of the production function defined by (18), as the marginal rate of substitution is equal to $\omega = \beta k^{1/\sigma} Y^{\delta}$ and $\log k$ is equal to

$$\log k = -\sigma \log \beta + \sigma \log \omega - \sigma\delta \log Y$$

We shall complete this section by briefly touching on the straightforward generalization of the *NH-CES* production functions to the n-factors case. It is generally expressed as

$$\sum_{i=1}^{n} C_i(Y) X_i = 1 \tag{19}$$

where $X_i = x_i^{-\rho}$ for $\sigma \neq 1$ or $X_i = \log x_i$ for $\sigma = 1$ $(i = 1, \ldots, n)$ and $C_i(Y)$ are all monotone functions of Y. The almost-homogeneous family is, thus, equal to

$$\sum_{i=1}^{n} \beta_i Y^{\gamma_i} x_i^{-\rho} = 1 \tag{20}$$

and the separable family is

$$Y = \frac{\sum_{i=1}^{k} \beta_i x_i^{-\rho} + \theta_1}{\sum_{j=k+1}^{n} \beta_j x_j^{-\rho} + \theta_2} \tag{21}$$

It is, of course, possible to have what may be called "block" homothetic or "block" non-homothetic *CES* functions depending upon the specific forms of $C_i(Y)$. For instance, if $\partial C_i/\partial Y \equiv 0$ for $i = 1, \ldots, k$ and $\partial C_i/\partial Y \neq 0$ for $i = k + 1, \ldots, n$, we will then have a k-block homothetic and $(n - k)$ block non-homothetic *CES* function,

$$X_1 + \cdots + X_k + C_{k+1}X_{k+1} + \cdots + C_n X_n = 1 \tag{22}$$

In the above formulation the marginal rates of substitution between x_i and x_j are independent of the level of output whenever x_i and x_j belong to the homothetic block $(i,j = 1, \ldots, k)$, but they are dependent on the output level whenever x_i and x_j belong to the non-homothetic block $(i,j = k + 1, \ldots, n)$.

III. Empirical Applications

An attempt was made to estimate the coefficients of the marginal rate of substitution function underlying the class of *NH-CES* production functions. In particular, equation (2) was used to estimate the iso-*MRS* family of *CES* functions with a constant non-homotheticity parameter. The primary objective of the empirical study was to determine whether or not the non-homothetic *CES* production function is more realistic than the ordinary *CES* function. The log-linear least squares estimation technique was applied to time-series data (1949–65) for two-digit *U.S.* manufacturing industries.[6]

Three models were tested: In Model I the underlying production function is of the ordinary *CES* (*H-CES*) type with Hicks-neutral technical progress; Model II as-

[6]Sources: Internal Revenue Service, U.S. Bureau of the Census, and U.S. Office of Business Economics

sumes the *H-CES* function with factor-augmenting technical progress; and in Model III the production function is of the iso-*MRS* non-homothetic *CES* type with a constant non-homotheticity parameter. The technical progress function in Model III is either the Hicks neutral or the factor-augmenting type. The estimated marginal rate of substitution function for each model and the underlying production functions are:

Model I:
Production Function

$$Y = A(0)e^{\epsilon t}[\beta L^{-\rho} + (1-\beta)K^{-\rho}]^{c}$$

Estimated Equation

$$\log k = a_1 + a_2 \log \omega + u \tag{23}$$

Model II:
Production Function

$$Y = A(0)e^{\epsilon t}[\beta \bar{L}^{-\rho} + (1-\beta)\bar{K}^{-\rho}]^{c}$$
$$\bar{L} = A_1(0)e^{\epsilon_1 t}L$$
$$\bar{K} = A_2(0)e^{\epsilon_2 t}K$$

Estimated Equation

$$\log k = a_1 + a_2 \log \omega + a_4 t + u \tag{24}$$

Model III:
Production Function

$$\bar{K}^{-\rho} + \frac{\beta}{1-\beta}(A(0)e^{\epsilon t}Y)^{\delta}\bar{L}^{-\rho} = H(A(0)e^{\epsilon t}Y)$$
$$\bar{L} = A_1(0)e^{\epsilon_1 t}L$$
$$\bar{K} = A_2(0)e^{\epsilon_2 t}K$$

Estimated Equation

$$\log k = a_1 + a_2 \log \omega + a_3 \log Y + a_4 t + u \tag{25}$$

In these estimated equations,[7] the coefficients are equal to:

$a_1 = \log[(\beta/1-\beta)^{-\sigma}C]$ (the distribution parameter), where

$$C = C[A(0), A_1(0), A_2(0)]$$

$a_2 = \sigma$ (the elasticity of substitution)

$a_3 = -\delta/(1+\rho)$ (the non-homotheticity parameter)

$a_4 = (\epsilon_1 - \epsilon_2 - \delta\epsilon\sigma)$ (technical progress parameters)

The results of estimation are shown in Table 1. First, we compare Model I with Model III. (Model III can also represent the case of Hicks neutral with the non-homothetic *CES*.) In all cases, Model III (*NH-CES*) provides superior estimates to the ordinary *CES* assumption. This is not surprising, as the *NH-CES* case contains two additional variables Y and t. The factor-augmenting hypothesis (Model II) represents a major improvement over the *H-CES* case. However, *it is very interesting that the NH-CES case with factor-augmenting technical progress gives more meaningful estimates in terms of both T-values and* R^2 *for twelve (out of nineteen) industries.* (The asterisk notation * is given to such industries in Table 1.) Hence, for *a majority of the industries, the non-homotheticity parameter is not equal to zero.* Non-homothetic *CES* functions prove to be a more realistic assumption than the ordinary *CES* case.

The elasticity of substitution is, of course, not the same for all industries. However, in a majority of the industries studied, statistically significant values not too far from unity were obtained with the exception of autos where the elasticity is 12.20 and the T-value of 2.28. In the twelve industries where the non-homothetic *CES* functions give superior fits, *the elasticity of substitution tends to be smaller than that of the homothetic case.* For instance, in tobacco, the elasticity of substitution is 1.13 with the non-homothetic parameter and is 2.07 without it. The tendency of the elasticity of substitution to decline when eliminated from the non-homothetic function can be attributed to decomposition into a "pure" substitution effect and an "expansion effect."

Although all of the essential parameters

[7] In the case of Model III, one is aware, of course, of the multicollinearity between Y and t. This makes it difficult to distinguish between Models II and III. Perhaps cross-section data should be utilized for identifying the *NH* parameters.

TABLE 1—COMPARISON BETWEEN *H-CES* AND *NH-CES* PRODUCTION FUNCTIONS

		Distribution Parameter	Substitution Parameter		Non-Homotheticity Parameter		Technical Progress Parameter			
Industry	Model	a_1	$a_2 = \sigma$	*T*-value	a_3	*T*-value	a_4	*T*-value	*F*-value	R^2
21: Tobacco	I	−1.078	2.07	7.228					52.24	0.7769
	II	−1.069	1.24	14.32			−0.0245	−7.043	135.6	0.9509
	III*	12.64	1.13	26.23	1.002	−5.761	0.0124	1.855	309.3	0.9862
22: Textile	I	−0.3930	0.487	2.833					8.024	0.3485
	II	0.4951	0.768	7.213			−4.833	−1.568	1.924	0.9649
	III	6.3202	0.836	6.609	−0.376	−1.686	−0.0411	−7.948	146.08	0.9712
23: Apparel	I	0.763	0.650	2.671					7.132	0.3222
	II	0.831	0.765	9.644			−0.034	−16 126	195.18	0.9654
	III	10.374	0.725	10.912	−0.621	−0.296	−0.001	−0.817	163.26	0.9741
24: Lumber	I	0.432	1.325	5.01					25.10	0.626
	II	−0.014	0.612	9.655			−0.035	−5.956	59.14	0.894
	III	5.112	0.655	8.531	−0.339	−1.430	−0.026	−3.111	43.05	0.908
25: Furniture	I	0.371	6.02	0.102					0.01	0.0007
	II	0.352	5.32	0.107			0.002	0.0457	0.006	0.008
	III	−42.330	1.38	0.382	2.96	0.807	−0.093	−0 747	0.221	0.049
26: Paper and Pulp	I	−0.192	1.16	9.978					99.55	0.869
	II	−1.292	0.62	7.605			−0.028	−3.712	99 08	0.934
	III*	6.654	0 67	7.595	−0.512	−2.161	−0.001	−0.099	84.92	0.951
27: Printing	I	7.290	0 89	4.921					24.21	0.617
	II	5.046	1.40	6.976			−0.021	−8.96	116.2	0.943
	III*	20.73	1.23	9.366	−0.974	−3.087	0.021	1.55	127.8	0.967
28: Chemicals	I	1.087	23.25	2.06					4.24	0.220
	II	1 031	41.67	0.823			−0.005	−0 871	2.465	0.260
	III	11.161	33.3	1.04	−0.642	−1 497	0 038	1 303	2.535	0 369
29: Petroleum	I	3.076	10.31	0.682					0.465	0.03
	II	7.843	1.172	2.462			−0 048	−2 34	3.04	0 303
	III*	28.449	1.008	5.899	−1 348	−6.871	−0.011	−1 007	24.452	0 849
30: Rubber and Plastics	I	0.454	6.62	1.27					1.61	0.097
	II	0.325	2 11	2.102			−0.014	−1 653	2.266	0.245
	III*	10.562	1.75	3.895	−0.725	−4.548	0.041	3.094	10.53	0.708
31: Leather	I	1 181	0.412	3.746					14.036	0.483
	II	0.763	0.735	4.05			−0.03	−7.39	59.40	0.895
	III*	11.40	0.825	7.708	−0.754	−7.211	−0.014	−4.75	201.17	0.979
32: Stone, Clay, Glass	I	0.106	1.59	13.11					171.97	0.919
	II	−0.188	0.991	7.695			−0.023	−3.042	137.90	0 952
	III	5.438	1.075	7 157	−0.369	−1.764	−0.001	−0.095	106.83	0 961
33: Primary Metal	I	6.247	1.11	11.41					130.23	0.897
	II	8.924	0.725	9.33			−0 03	−3.533	121.2	0 945
	III	14.915	0.882	5.995	−0.456	−1.87	−0.004	−0.256	96.43	0 957
34. Fabricated Metal	I	2.014	4.525	0.952					0.907	0.057
	II	7.960	0.870	5.75			−0.03	−6.01	19.63	0 737
	III*	19.876	2.079	1.231	−1.031	−1.932	0.033	0.995	16.88	0 796
35: Machinery	I	5.811	1.181	7.489					56 09	0 789
	II	7.739	0.873	5.752			−0.014	−1.765	33.56	0.827
	III*	11.981	1.264	3.203	−0.397	−2.08	0.005	0.474	29.16	0.870
36: Electrical	I	4.098	1.618	4.49					20.20	0.574
	II	5 549	1.199	3.514			−0.008	−1.113	10.88	0.61
	III*	14.761	3.175	0.975	−0.812	−2.13	0.063	1.865	10.59	0.71
37: Autos	I	0.047	10.20	0.747					0.559	0.036
	II	0.12	10.53	0.704			−0.006	−0.459	0.37	0.050
	III*	15.615	12.20	2.28	−1.005	−13.655	0.052	9.543	65.67	0.938
38: Instruments	I	0.291	9.71	1.209					1.463	0.088
	II	0.179	2.12	2.65			−0.02	−2.285	3.549	0.336
	III*	10.875	3.45	2.14	−0.739	−3.802	0.031	2.079	9.458	0.686
39: Miscellaneous	I	0.060	2.41	4.540					20.61	0.579
	II	0.204	1.176	6.738			−0.030	−4.012	28.72	0.804
	III*	−5.338	0.915	7.640	0.374	2.553	−0.06	−4.50	28.87	0.869

of the non-homothetic *CES* function are estimated, there is still a specification problem regarding the $H(Y)$ function. This function may be ascertained from a second step of estimation. The left-hand side of the equation

$$\bar{K}^{-\rho} + \frac{\beta}{1-\beta}(A(0)e^{u}Y)^{\delta}\bar{L}^{-\rho} = H(A(0)e^{u}Y)$$

can be computed from the estimated values of the parameters for given values of $Y(t)$, $\bar{K}(t)$, and $\bar{L}(t)$. Using again the left-hand side H may be estimated as some function of $A(0)e^{u}Y$. For instance, it may take a linear or a log-linear form. Only after H is known, can the specific type of *NH-CES* functions be identified (such as separable or almost-homothetic, etc.). While this second step estimation has not been done for the purpose of this paper, it is of no consequence for two reasons. First, the estimation method outlined is *ad hoc* and there is no econometric method available that would guarantee the desirable estimation results. Second, and more importantly, the purpose of this paper is not to identify specific forms of *NH-CES* functions, but to compare the classes of *CES*, homothetic and non-homothetic (in whatever form) production functions, and to present preliminary empirical results which suggest that one class (*NH-CES*) is more realistic and meaningful than the other (*H-CES*).

REFERENCES

K. Arrow et al., "Capital-Labor Substitution and Economic Efficiency," *Rev. Econ. Statist.*, Aug. 1961, *43*, 225–50.

F. W. McElroy, "Note on the CES Production Function," *Econometrica*, Jan. 1967, *35*, 154–56.

D. McFadden, "Constant Elasticity of Substitution Production Functions," *Rev. Econ. Stud.*, June 1963, *30*, 73–83.

J. Paroush, "A Note on the CES Production Function," *Econometrica*, Jan.–Apr. 1964, *32*, 213–14.

P. A. Samuelson, "Using Full Duality to Show That Simultaneously Additive Direct and Indirect Utilities Implies Unitary Price Elasticity of Demand," *Econometrica*, Oct. 1965, *33*, 781–96.

R. Sato, "The Estimation of Biased Technical Progress and the Production Function," *Int. Econ. Rev.*, June 1970, *11*, 179–208.

——, "On the Class of Separable Non-Homothetic CES Functions," *Econ. Stud. Quart.*, Apr. 1974, *15*, 42–55.

——, (1975a) "The Most General Class of CES Functions," *Econometrica*, July–Sept. 1975, *43*, 999–1003.

——, (1975b) "On Homothetic and Holothetic Production Functions," presented at the 5th World Congress of the Econometric Society, Toronto, Aug. 1975.

——, (1975c) "Analysis of Production Functions by Lie Theory of Transformation Groups: Classification of General CES Functions," presented at the Toba Symposium on Economic Theory, Dec. 1975 (Proceedings, forthcoming in 1977).

——, "On Self-Dual Preferences," *Econometrica*, Sept. 1976, *44*, 1017–32.

Internal Revenue Service, *Statistics of Income*; *Corporate Income Tax Returns*, Washington 1949–65.

U.S. Bureau of the Census, *U.S. Census of Manufactures*, Washington 1949–65.

U.S. Office of Business Economics, *Surv. Curr. Bus.*, Washington 1949–65.

Who Benefits from Economic Development?—
A Reexamination of Brazilian Growth in the 1960's

By GARY S. FIELDS*

One of the most interesting and controversial cases of recent economic development is that of Brazil. Over the decade of the 1960's, Brazil achieved a substantial rate of growth by the standards of less developed countries (*LDC*). For the latter years of the 1960's and the first part of the 1970's, growth rates approached 10 percent per annum. On this basis, the Brazilian case was widely heralded as an "economic miracle."

More recently, however, two challenges have arisen. One group of analysts has looked with disfavor upon social policies which prevailed over the period, particularly following the rise to power in 1964 of the military government. A second group examined the distributional question of who received the benefits of this growth, found greater income inequality according to conventional measures, and concluded that the poor benefited very little if at all. These observations have caused many students of development to ask whether the high rate of aggregate growth in Brazil was worth the apparent social and distributional costs. The consequent debate, involving Albert Fishlow (1972, 1973a,b), Carlos Langoni (1972, 1975a,b), and Celso Furtado, among others, has been intense and often acrimonious, resulting in widespread disagreement about the desirability of taking Brazilian economic and social policies as a model for other developing countries to follow.

The purpose of this paper is to reexamine one of these two challenges, namely, the distributional impact of Brazilian economic growth during the 1960's. My results lead to a quite different interpretation from the conventional one. I will show that the poor in Brazil *did* participate in the rapid economic growth of the decade. Estimates presented below indicate that average real incomes among families defined as poor by Brazilian standards increased by as much as 60 percent while the comparable figure for nonpoor families is around 25 percent. However, since nonpoor families receive incomes which are much greater than those of poor families, the bulk of the growth of national income over the decade was received by families whose incomes placed them above the official poverty standard. Thus, it would be *incorrect* to say either that 1) in achieving a high rate of economic growth in Brazil the rich got absolutely richer while the poor got absolutely poorer, or 2) the incomes of poor families increased more slowly (percentagewise) than those of nonpoor families. These and other findings are presented below in Section II, and some of the reasons for the observed changes are discussed in Section III.

In assessing the distributional consequences of Brazilian economic growth, this study explicitly adopts an *absolute* poverty approach. In so doing, it is at odds with the bulk of the economic development literature, which while urging a poverty focus, has long relied on measures of *relative* income inequality and Lorenz curves. Thus, this paper does not merely offer "one more measure"; it is, rather, the use of a different *type* of measure that causes the divergent

*Associate professor of economics, Yale University. An earlier draft of this paper was written while I was a visiting professor at the Centro de Estudios sobre Desarrollo Económico, Universidad de Los Andes, Bogotá, Colombia. Partial support for this research was received from the International Bank for Reconstruction and Development under RPO/284. However, the views expressed do not necessarily reflect those of IBRD. I wish to thank the above institutions without implicating them.

TABLE 1—BRAZILIAN SIZE DISTRIBUTION OF INCOME AND ECONOMICALLY ACTIVE POPULATION, 1960 AND 1970

A. Variable Income Brackets

Monthly Income in 1960 NCr$[c]	Percentage of Population	Percentage of Income	Monthly Income in 1970 NCr$[c]	Percentage of Population	Percentage of Income
None	14.7	0.0	None	11.7	0.0
0–2.1	22.3	5.2	1–100	31.7	8.0
2.1–3.3	14.4	7.0	101–150	12.8	6.2
3.3–4.5	10.5	7.4	151–200	15.6	10.6
4.5–6.0	13.1	12.3	201–250	4.5	3.9
6.0–10.0	13.8	20.0	251–500	14.6	21.2
10.0–20.0	8.2	22.2	501–1000	5.9	17.1
20.0–50.0	2.6	16.4	1001–2000	2.2	13.0
Over 50.0	0.5	9.4	2001 and over	1.0	20.1
Mean (Current NCr$)	5.52		Mean (Current NCr$)	258.1	
Mean (1960 *U.S.* $ per year)	513		Mean (1960 *U.S.* $ per year)	679	
Gini Coefficient	0.59		Gini Coefficient	0.63	

B. Comparable Income Brackets[a]

Monthly Income in 1960 NCr$[c]	Percentage of Population		Cumulative Percentage of Population	
	1960	1970[b]	1960	1970[b]
None	14.7	11.7	14.7	11.7
0–2.1	22.3	23.8	37.0	35.5
2.1–3.3	14.4	12.2	51.4	47.7
3.3–4.5	10.5	11.0	61.9	58.6
4.5–6.0	13.1	14.5	75.0	73.1
6.0–10.0	13.8	9.4	88.8	82.5
10.0–20.0	8.2	10.9	97.0	93.4
20.0–50.0	2.6	5.0	99.6	98.4
Over 50.0	0.5	1.6	100.1	100.0

Source: Panel A, Albert Fishlow (1972), Tables 1 and 5
[a]Calculated from data in Panel A.
[b]Approximations.
[c]NCr$ (thousands).

results. The paper concludes in Section IV by reviewing the principal findings and exploring some further questions of more general applicability raised by the Brazilian debate.

I. Basic Results and the Customary Interpretation

The best known study of economic growth and changes in the size distribution of income in Brazil over the decade of the 1960's is that of Fishlow (1972). The basic data are reported in Panel A of Table 1. Looking first at the level of income, the mean income among the economically active population in constant *U.S.* dollars increased from $513 in 1969 to $679 in 1970, a real increase of 32 percent.[1]

At first glance, however, the data on income distribution seem to tell another story. We see that the upper 3.2 percent of the economically active population received 27 percent of the income in 1960; by 1970, their share had risen to more than 32 percent. In addition, the Gini coefficient rose from 0.59 to 0.63, seemingly implying a less

[1]This is the percentage increase of "uncorrected incomes" for the "total economically active population," the only comparison possible with Fishlow's data.

even income distribution. A second study of Brazilian growth over the same period, by Langoni (1972, 1975a), arrives at basically the same changes in the income distribution.[2]

In research on the distributional consequences of economic development, virtually all studies to date have maintained (usually implicitly) that changes in economic well-being are positively related to changes in the *level* of national income and negatively related to changes in measured *inequality* in its distribution, using such measures as the Gini coefficient or the share of income accruing to the poorest 40 percent. In accordance with this type of judgment, Fishlow's interpretation of the rising Gini coefficient and income share of the very richest is the following: "The conclusion that inequality has increased over the course of the decade accordingly seems correct, if lamentable" (1972, p. 399). The qualitative result—of a "worsening" income distribution in Brazil—has been widely accepted.[3,4]

Contrary to the customary interpretation, a reexamination of the Brazilian data from an absolute income perspective tells a different story. This is the subject of Section II.

II. A Reexamination

A. *The Distribution of the Benefits of Growth*

The fundamental question underlying the analysis of income distribution in economic development is this: who (as classified by income class or other economic or socioeconomic criterion) receives what share of the proceeds of economic growth? The ideal way to answer this question would be to follow the same set of individuals over a period of time to see how their incomes change, and how these changes relate to their initial income position and other characteristics. The type of longitudinal (or panel) data needed to do this do not exist for Brazil. In their absence, we must rely on frequency distributions of the population by income class.

To measure changing absolute incomes, the numbers presented in Table 1 do not quite suffice, because they have different income brackets in the two years.[5] Lacking the raw data with which to make an exact fit, it is necessary to take the income brackets for one year as a base and to approximate the frequency from the other year in each category. The actual distribution for 1960 and the approximate values for 1970 are shown in Panel B of Table 2.[6,7]

[2]Using slightly different definitions than Fishlow and excluding unremunerated workers and the unemployed, Langoni found a rise in the Gini coefficient from 0.49 to 0.56, a falling share of national income received by each of the four lowest quintiles, and a rising share received by the richest 5 percent (from 27.9 to 34.9 percent of national income). Langoni's exclusion of zero-income persons is presumably the reason why his Gini coefficients are lower than Fishlow's. For an English-language description detailing the characteristics of these and other income distribution sources, see Langoni (1975b).

[3]See, for instance, the work of Irma Adelman and Cynthia Taft Morris, William Cline, and Adolfo Figueroa and Richard Weisskoff

[4]The Brazil debate has been conducted largely on the basis of the undisputed rise in the Gini coefficient However, it is well known that when Lorenz curves cross, as they have been shown to do in Brazil, some indices of relative income inequality may indicate a more equal distribution of income, while others may indicate the reverse. (For an empirical illustration of this point for three Latin American countries, see Weisskoff.) Possible ambiguities in Brazil were apparently put to rest by Langoni (1972, p. 15), who reported increases in the variance of logarithms and the Theil inequality index as well as the Gini coefficient. More recently, though, Samuel Morley and Jeffrey Williamson reported that the relative inequality measure proposed by A. B. Atkinson yields the opposite result.

[5]This is also true for other sources of income distribution statistics; see Langoni (1975b).

[6]The procedure used to approximate the 1970 distribution is the following. The mean incomes in 1960 and 1970 were $513 and $679, respectively, both measured in constant 1960 *U.S.* dollars. These same means, expressed in current new cruzeiros (NCr$), were 5.52 and 258.1. Thus, the ratio of the real means was 1.32, and of the nominal means 46.76. The ratio of these, 35.32, is then an inflation factor which can be used to deflate the 1970 brackets. For example, the first positive income bracket in 1970 runs from 0 to 2.8 constant NCr$. Then applying a linear approximation to the population frequency within each bracket, 2.1/2.8 of the population in the 0–2.8 category was assigned to the 0–2.1 category, and the remaining 0.7/2.8 was assigned to the next higher category. An analogous procedure was followed for the other brackets. It would, of course, have been better to have

The most striking feature of these data is that the cumulative percentage of population was lower in 1970 than in 1960 for *every* income bracket. This means, very simply, that the economic growth which took place over the decade reached persons at all income levels, and not just those at the top.

It should be observed that these figures refer to percentage of the population. With a growing population, these data imply that the Brazilian economy was able to create opportunities for its economically active population to earn higher incomes at a faster rate than its labor force was expanding.

These findings clearly refute the notion that the rich got absolutely richer while the poor got absolutely poorer in Brazil during the 1960's.

B. *Income Growth of the Poor and Nonpoor*

The analysis may be extended to compare the income growth of the poorest groups with that of all others. We may ask four related questions:

1) Defining "the poor" as those whose incomes were below a constant real poverty line, did the fraction of the economically active population classified as poor increase or decrease over the decade, i.e., was the incidence of poverty being reduced?

2) What were the rates of increase of income among the poor and nonpoor, i.e., were the remaining poor getting less poor, absolutely and relatively?

3) How much of the economic growth over the decade went to the poor and how much to the nonpoor?

4) Defining the "poverty gap" as the amount by which poor persons' incomes would have to be raised to bring them all up to the poverty line, how much of the gap was filled during the decade?

We must begin by establishing a poverty line. Something like 31 percent of Brazilian families were poor in 1960 by Brazilian definitions.[8] Since it is not possible to identify these families exactly, we may suppose that those persons in the two lowest income brackets (i.e., less than 2.1 NCr$ constant), which in 1960 comprised 37.0 percent of the population, were below the poverty line. From now on, we will refer to these persons with incomes below 2.1 as "the poor" and the rest of the population as "the nonpoor."

Considering first the question of changing numbers of poor, we see from Panel B that there was a small decrease in the percentage of the economically active population with incomes below the poverty line, from 37.0 to 35.5 percent. There was not a higher incidence of poverty in 1970 than in

---

used the exact distribution of the economically active population across these income categories rather than this approximation; but owing to the lack of a public use sample for the microeconomic data, this was impossible.

One may ask whether the simplified linear interpolation introduces a bias into Panel B and subsequent calculations, and if so how great that bias is. The answer is that the income share of those in the 0–2.1 income class is overstated by my assumptions, but under *no* possible alternative assumptions would any of the conclusions reached below, in particular, the conclusion about the relative rates of income growth among the poor and nonpoor, be reversed qualitatively. Details of these calculations may be found in the author (1976).

[7]A referee has noted that the data used in this paper include zero-income persons. He argues that these were previously unpaid farm workers who were forced off the land during the 1960's and took up wage employment. If this is correct, it raises the possibility that the income gains observed in the statistics are more apparent than real (due to the receipt of cash incomes, which are recorded, rather than incomes-in-kind, which were not). To resolve this doubt, he suggested excluding the zero-income labor force and reestimating income levels at the lower end of the distribution. I performed these additional estimates and found that the conclusions presented below are sustained. The calculations are available upon request.

[8]The poverty line is defined according to Brazilian standards. Says Fishlow. "The real minimum wage for 1960 in the Northeast, the poorest region, is taken as the lower limit of acceptable income for a family of 4.3 persons. For rural Brazil, the wage prevailing in the rural areas of the Northeast is taken; for the urban Northeast, the standard of medium sized municipio is applied; and for all other urban residents, the Northeast level, incremented by 15 percent to allow for higher relative prices, is applied. The poverty line for different size families is defined with the aid of the elasticity of expenditure on food with respect to family size; because of economies of scale larger families need relatively less income, and conversely for smaller" (1972, pp. 393–94).

TABLE 2—ANALYSIS OF ECONOMIC GROWTH IN BRAZIL AND THE UNITED STATES DURING THE 1960'S

Effect	Definition of Effect	Importance in the Economic Growth of:[a] Brazil 1960–70	U.S. 1959–69
α = Enlargement of the higher income sector = Change in the number of persons in the high income sector, multiplied by the income differential between the high income and low income sectors in the base year;	$(f_n^{70} - f_n^{60})(\bar{y}_n^{60} - \bar{y}_p^{60})$	6	19
β = Enrichment of the high income sector = Change in income within the high income sector, multiplied by the number *of persons in that sector in the base year;*	$(\bar{y}_n^{70} - \bar{y}_n^{60}) f_n^{60}$	82	72
γ = Interaction between enlargement and enrichment of the high income sector = Change in income within the high income sector, multiplied by the change in the number of persons in that sector;	$(\bar{y}_n^{70} - \bar{y}_n^{60})(f_n^{70} - f_n^{60})$	2	8
δ = Enrichment of the low income sector = Change in income within the low income sector, multiplied by the number of persons in that sector in the terminal year;	$(\bar{y}_p^{70} - \bar{y}_p^{60}) f_p^{70}$	10	1
$\alpha + \delta$ = Sum of "poor" enlargement and enrichment effect		16	20
Total		100	100

Sources: Brazil (see text); *U.S.*: *Statistical Abstract of the United States 1971*, Tables 485, 512, 513, 515, 517.
Notes: f_p = fraction of the population which was poor.
f_n = fraction of the population which was nonpoor.
$\bar{y}_p$ = average income of the poor population.
$\bar{y}_n$ = average income of the nonpoor population.

Brazil 1960	Brazil 1970	United States 1959	United States 1969
f_p^{60} = 37.0%	f_p^{70} = 35.5%	f_p^{59} = 23.8%	f_p^{69} = 14.9%
f_n^{60} = 63.0%	f_n^{70} = 64.5%	f_n^{59} = 76.2%	f_n^{69} = 85.1%
$\bar{y}_p^{60}$ = NCr\$0.8	$\bar{y}_p^{70}$ = NCr\$1.3	$\bar{y}_p^{59}$ = U.S.\$2,423	$\bar{y}_p^{69}$ = U.S.\$2,689
$\bar{y}_n^{60}$ = NCr\$8.3	$\bar{y}_n^{70}$ = NCr\$10.6	$\bar{y}_n^{59}$ = U.S.\$10,774	$\bar{y}_n^{69}$ = U.S.\$12,343

[a]Shown in percent.

1960, as might have been supposed from the rising inequality coefficients. Neither, though, was the incidence of poverty substantially reduced.

Next, let us compare the rates of growth of incomes among the poor as opposed to the nonpoor. To determine the average income in each group in each year, we use the basic accounting identity that total income, in the economically active population as a whole or for the poor and nonpoor subpopulations, is equal to the mean income multiplied by the number of persons in question. Letting $\bar{y}_p$ and $\bar{y}_n$ be the mean incomes of the poor and nonpoor, respectively, and P be the population, we have, for 1960,

$$(1) \quad 37.0\% P^{60}\bar{y}_p^{60} + 63.0\% P^{60}\bar{y}_n^{60} = P^{60}(5.52)$$

(2) $37.0\% P^{60}\bar{y}_p^{60} = 5.2\% P^{60}(5.52)$

and for 1970,

(3) $35.5\% P^{70}\bar{y}_p^{70} + 64.5\% P^{70}\bar{y}_n^{70} = P^{70}(258.1/35.32)$

(4) $35.5\% P^{70}\bar{y}_p^{70} = 8.0\% P^{70}(2.1/2.8)(258.1/35.32)$

Equations (1) and (3) tell us simply that total income is equal to the sum of the incomes of the poor and nonpoor. In equations (2) and (4), the incomes of the poor are expressed first as their mean income multiplied by the number poor, and then as income in the economically active population multiplied by the income share of the poor. The numbers in parentheses are explained in footnote 6.

Solving, we find for the poor:

(5) $\bar{y}_p^{60} = 0.8 \qquad \bar{y}_p^{70} = 1.3$

$$\frac{\bar{y}_p^{70} - \bar{y}_p^{60}}{\bar{y}_p^{60}} = 63\%$$

and for the nonpoor:

(6) $\bar{y}_n^{60} = 8.3 \qquad \bar{y}_n^{70} = 10.6$

$$\frac{\bar{y}_n^{70} - \bar{y}_n^{60}}{\bar{y}_p^{70}} = 28\%$$

From (5), we see that the poor became noticeably less poor; their incomes are estimated to have grown by 63 percent over the decade. Furthermore, comparing (5) and (6), the incomes of the poor appear to have grown at a rate more than double that of the nonpoor.[9] This reinforces the earlier observation that the rich in Brazil did not benefit during the 1960's at the expense of the poor.

Is the 1970 distribution of incomes between poor and nonpoor more or less equal than the 1960 distribution? The answer depends on how one defines "equal." If absolute real income differentials are our standard, we observe

(7) $\bar{y}_n^{60} - \bar{y}_p^{60} = 7.5 \qquad \bar{y}_n^{70} - \bar{y}_p^{70} = 9.3$

and see that the absolute gap widened by about 25 percent. However, this gap was a smaller percentage of per capita income in 1970 than in 1960:

(8)
$$\frac{\bar{y}_n^{60} - \bar{y}_p^{60}}{\bar{y}^{60}} = \frac{7.5}{5.52} = 1.36$$
$$\frac{\bar{y}_n^{70} - \bar{y}_p^{70}}{\bar{y}^{70}} = \frac{9.3}{258.1/35.32} = 1.27$$

Furthermore, if we take relative income ratios as our standard for comparison, we find

(9) $\bar{y}_n^{60}/\bar{y}_p^{60} = 10.4 \qquad \bar{y}_n^{70}/\bar{y}_p^{70} = 8.2$

that is, a reduction of the ratio of nonpoor to poor incomes of about 20 percent. The interpretation of these figures is a matter of individual judgment.

Now let us address the question of how much of the economic growth over the decade went to the poor and how much to the nonpoor. In my 1975 paper, I have devised a methodology for decomposing total economic growth into four effects pertaining to the enlargement of the various sectors and the enrichment of persons within them. The specific formulas, and the numerical results for Brazilian economic growth during the 1960's, are given in Table 2. The outstanding result is that the bulk of economic growth in Brazil accrued to persons who had been above the poverty line in 1960 (β = 82 percent). Of the total

[9]The specific figure is open to question for two offsetting reasons. On the one hand, income growth of the poorest 37.0 percent is understated, since some 4 percent of the poor (1.5%/37.0%) received large enough income increases to raise them above the poverty line, and their incomes appear as nonpoor incomes in the above calculations. On the other hand, the average income of the poor in 1970 ($\bar{y}_p^{70}$) tends to be overstated, owing to the fact that although the poorest of the poor received incomes below the average for their income category, they were assigned the average value in the approximations of Table 1 and equation (4). Notwithstanding these doubts, it is certain that the incomes of the poor grew at least as rapidly as those of the nonpoor, since it is mathematically impossible for the data to be consistent with the alternative hypothesis (incomes of the poor growing more slowly); see fn. 7. At issue is *how much* greater was the increase for the poor. Although we work with the data in Table 1 in what follows, the reader should remember that these are only approximate values and not exact figures.

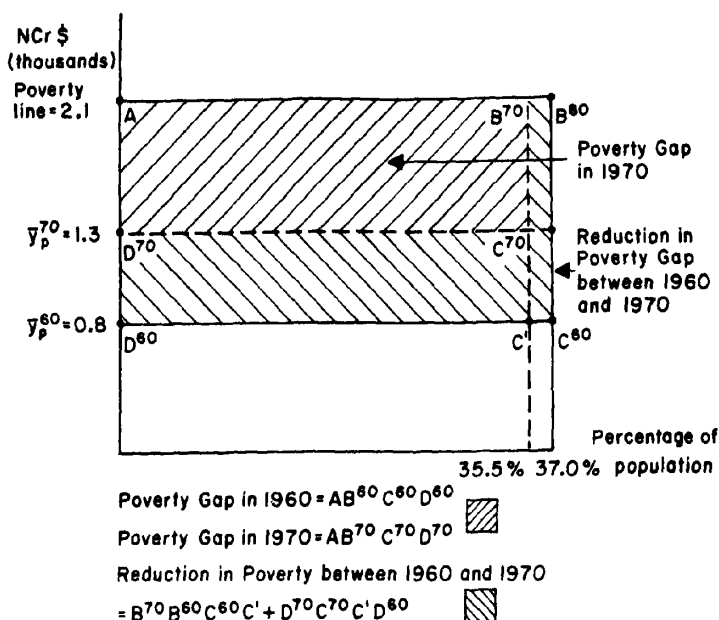



FIGURE 1. POVERTY GAP IN BRAZIL, 1960 AND 1970

growth, only about 16 percent ($\alpha + \delta$) went to the poor. Of this, 6 percentage points went to elevating formerly poor persons above the poverty line (α) while the other 10 percentage points served to make the poor somewhat less poor (δ).

In interpreting this pattern, two considerations should be borne in mind. For one thing, it is not really surprising that most of the economic growth of a country would be received by the nonpoor. This is partly because higher income persons have superior access to income-earning opportunities; partly because many countries develop by creating more employment of professional and skilled workers, who are likely to have been earning above the poverty line to begin with; and partly because of the simple mathematical fact that the poor cannot receive a very large share of the income growth before they are no longer poor. In addition, if we compare the percentage of growth accruing to the poor in Brazil (16 percent) with the same figure for the United States for the same decade (20 percent), we find that the results are not very different, despite the reputation of the United States as a relatively more egalitarian society. In future research, it would be most interesting to compare α, β, and δ in a number of *LDCs* and to try to understand similarities and differences in the observed patterns.

Finally, we may examine the extent to which the Brazilian economy closed its poverty gap during the 1960's. The poverty gap is calculated as the sum of the differences between each poor person's (or family's) income and the poverty line. This concept may be illustrated with the aid of Figure 1. Poor persons in 1960, who comprised 37.0 percent of the population (P), received an average income of NCr\$0.8. The poverty gap then was:

(10) Poverty gap in 1960

= (Poverty line minus mean income of persons below the line in 1960) x (Population below the poverty line)

= (\$2.1 − \$0.8) x 37.0% P

= \$48.1% P

The poverty gap in 1960 is illustrated by the area $AB^{60}C^{60}D^{60}$. Similarly, for 1970, we have

(11) Poverty gap in 1970

$= (\$2.1 - \$1.3) \times 35.5\% P$

$= \$28.4\% P$

given by area $AB^{70}C^{70}D^{70}$. Expressed as a percentage of population, the amount of the poverty gap made up during the 1960's is the sum of two components: that part of the increase in incomes which elevated some of the poor up to the poverty line ($B^{70}B^{60}C^{60}C'$), plus the increase in incomes of those who remained below the line ($D^{70}C^{70}C'D^{60}$). For Brazil between 1960 and 1970, the amount made-up was:

(12) Poverty gap made-up

= (Difference between poverty line and mean income of the poor in 1960) x (Number of poor elevated above the poverty line)

\+ (Change in mean income of the poor between 1960 and 1970) x (Number of poor remaining poor)

$= [(\$2.1 - \$0.8) \times 1.5\% P]$
$+ [(\$1.3 - \$0.8) \times 35.5\% P]$

$= \$19.9\% P$

The percentage of the poverty gap made-up in Brazil over the decade is the ratio of (12) to (10) or 41 percent.

Coincidentally, in the United States (see U.S. Bureau of the Census, Table 517) the poverty gap was reduced by exactly the same percentage (41 percent) over the same period, much of which comprised the "War on Poverty" years of the Johnson Administration. Although the percentage reduction was the same in the two countries, their patterns differed noticeably, as may be seen from the following data:[10]

	Brazil 1960-70	United States 1959-69
Percentage Reduction in Poverty Gap	41%	41%

[10]Sources: Brazil: computed from Table 1 in text; United States: U.S. Bureau of the Census, Tables 512, 513, 515, 517.

	Brazil 1960-70	United States 1959-69
Percentage Reduction in Fraction Poor	5%	33%
Percentage Reduction in Percentage Difference Between Average Income of the Poor and the Poverty Line	38%	20%
Fraction of Poverty Gap Reduction Attributable to Smaller Fraction of Population Below the Poverty Line	10%	61%

We observe that in Brazil, the poverty gap reduction took the form of substantially raising the incomes of the poor in percentage terms while elevating relatively few above the poverty line. In the United States, in contrast, the fraction poor was reduced by one-third, but those who remained poor were helped relatively less by a decade of growth than in Brazil. These observations support the view that poverty in Brazil involves individuals potentially in the mainstream of the economy (principally low-income workers) while poverty in the United States is often attributable to lack of economic activity (for example, among retirees and the physically and mentally handicapped).

C. *Summary*

Concerning the changes in income distribution in Brazil over the decade of the 1960's, this section has established the following:

1) The entire income distribution shifted in real terms, benefiting every income class.

2) There was a small decline in the fraction of the economically active population classified as below the poverty line, but those who remained poor received markedly higher incomes (in proportional terms).

3) The percentage increase in income for those below the poverty line was greater than the increase for those not in poverty, and may well have been twice as high, or more.

4) The relative income gap between poor and nonpoor persons narrowed in terms of ratios but widened absolutely.

5) The bulk of the income growth over

the decade accrued to persons above the poverty line. A similar pattern is observed for the United States, an allegedly more egalitarian society.

6) The poverty gap in Brazil was reduced by 41 percent between 1960 and 1970. The United States reduced its poverty gap by exactly the same percentage over the same decade.

III. How It Happened

How was the Brazilian economy able to shift its entire income distribution and eliminate a considerable percentage of its poverty gap during a decade of growth? The basic dimensions of change are given in Table 3.

For three-quarters of that country's economically active population, wages were the only source of income, and the income received by wage earners was 71 percent of the total. It follows, therefore, that the changing income distribution has its primary origin in a changing labor market.

Earnings are higher in urban than rural areas, and higher in industry than in agriculture. Thus, a shifting income distribu-

TABLE 3—SOME ASPECTS OF BRAZILIAN LABOR MARKETS IN THE 1960'S

Income Source, 1970[a]			
Wage earners as percentage of income recipients			74%
Income received by wage earners as percentage of total			71%
Median Earned Income By Rural-Urban, 1960 (approximate)[b]			
Urban and suburban households			Cr$1,250
Median Earned Income By Economic Sector, 1970 (approximate)[c]			
Industrial			NCr$195
Agriculture			110
All sectors			165
Population (in millions)[d]	1960	1970	Growth
Total	70.1	93.2	33%
Urban	32.5	52.1	60%
Rural	37.6	41.1	9%
Real Output By Sector, 1949 = 100[e]			
Industrial	261.4	511.8	96%
Agriculture	156.1	239.5	53%
Total real product	205.7	368.5	79%
Employment By Sector (in millions)[f]			
Industrial	3.0	5.8	77%
Agriculture	12.2	13.1	9%
Total economically active population	22.6	29.5	30%
Employment By Occupational Type (in thousands)[g]	1960	1969	
Primary: Agricultural activities, vegetable extraction, and fishing	12,271	12,533	2%
Secondary: Mineral extraction, industrial production and services, and construction	2,791	5,476	96%
Tertiary: Professionals, sellers of services (including repairmen and domestic workers), merchants, transport and communication workers, and civil servants (including police and army)	5,341	11,082	107%
Others not elsewhere classified	2,248	873	

[a]Comisión Economica para América Latina (1974), p. 22.
[b]Brasil (1960), Table 6.
[c]Brasil (1970), Table 8.
[d]Brasil (1960), Table 1 and Brasil (1970), Table 1.
[e]Fundação Getúlio Vargas (1973), Table 2.
[f]Brasil (1970), Table V.
[g]Singer (1971), Tables 2.V, 2.VI.

tion and reduction of poverty could result from the transfer of the population from rural agriculture to urban areas in general and to the industrial sector in particular. That is just what happened. The urban population grew nearly twice as fast as the total population and more than six times faster than the rural population, which can only be due to substantial rural-urban migration. Rates of growth of output and employment in the industrial sector were higher than in agriculture. The changing sectoral distribution of the labor force is reflected as well in the occupational distribution, the number of jobs at the lowest occupational levels increasing by just 2 percent over the decade, the number of jobs at higher levels doubling.

What caused labor market conditions to change as they did? The answer goes to the very heart of the Brazilian economic model. The main points of contention concern the role of government policy, particularly in four areas: industrialization and stabilization, international trade, government wage policy, and education. It is well beyond the scope of this paper to attempt to pass judgment on the relative merits of the opposing viewpoints. The interested reader is referred to Fishlow, Furtado, *Brazilian Trends*, Morley and Williamson, J. P. Wogart, Kenneth Mericle, John Wells, and Pedro Malan and Wells. How much of the improvement in labor market conditions is due to economic growth itself and how much to government policy remains an unsettled issue.

IV. Conclusion

A. *Recapitulation*

The conventional wisdom concerning Brazilian economic development over the 1960-70 period may be summarized by three propositions:

1) The absolute rate of growth was high, particularly in the latter part of the period.

2) Income distribution worsened.

3) Significant social and political costs were paid.

Many economists and other social scientists have invoked judgments on points 2) and 3) in questioning whether the higher rate of economic growth was "worth it." Evidence on the first two points is presented in Section I above.

Accepting the rapidity of aggregate growth over the decade, I have in this paper reexamined the challenge concerning the income distributional consequences of Brazilian growth. The main innovation is the use of *absolute poverty* measures in place of the usual *relative inequality* indices. Changes over the decade in absolute economic position of the poor and nonpoor populations were presented in Section II.

The findings based on the absolute poverty approach cast considerable doubt on the conventional wisdom. At minimum, the widely held notion that "the rich got rich at the expense of the poor" receives no support in the data examined here. To the contrary, the poor in Brazil clearly *did* share in a decade of economic development. Some poor were lifted out of poverty. For those left behind, their incomes grew at least as rapidly as those of the nonpoor. At the same time, the very rich also got richer than before, in both absolute and relative terms. Relative inequality did become greater by most measures.

Section III then described how changes in the structure of production and employment in the Brazilian economy shifted over the decade in favor of the relatively advanced and high-paying sectors: urban areas, the industrial sector, and relatively high-level occupations. These factors presumably account for a considerable part of the observed income distribution changes.

B. *Issues in Interpreting the Brazilian Experience*

In appraising the performance of the Brazilian economy over the 1960's, some important questions are raised by these findings:

1) *How much weight do we want to give in our evaluations to changes at which points on the income distribution?*

I have chosen in this paper to concentrate

on the number of very poor in Brazil and on the levels of income they receive. Such a focus has been urged by many writers, including Fishlow and Langoni themselves.[11] Nonetheless, the measures they use focus either on the entire income distribution or on the very top. In particular, it is well known (see, for example, David Champernowne) that the Gini coefficient assigns the greatest weight to changes in the *middle* of the income distribution and is relatively insensitive to income changes at either end. There thus appears to be a discrepancy between the welfare weights implicit in the measures used by previous writers on Brazil and the judgments they themselves wish to make about the primacy of income changes among the very poor.

2) *Are the incomes of the poor in Brazil being raised fast enough?*

An increase in average incomes of the poor of as much as 60 percent in a decade works out to an average annual rate of about 4 percent, starting from a very low level. I am unaware of evidence from other less developed countries that might indicate whether this rate is comparatively high or low. In any event, taking Brazil on its own, with a continuing growth rate of 5 percent per annum, 20 to 30 years would be needed to raise the poorest decile up to $100 per capita income, given the present pattern of income inequality. Writes Fishlow: "Can the present starving poor be expected to wait for 30 years, amid rising affluence, to attain the princely sum of $100 per capita? That, stripped of its niceties, is what the debate is all about" (1973, p. 90). Whether the growth experience of Brazil should be commended or condemned may well hinge on the answer to this question.

3) *Was the economic growth of the latter 1960's and early 1970's worth the apparent social and political costs?*

Very few studies of economic development have considered the noneconomic costs of growth. Yet, in the case of Brazil, this issue can hardly be avoided. In presenting these results on the distributional question, I have *not* taken a position in favor of the social measures adopted in Brazil, nor would I wish to. Conventional welfare economics offers no real guidance on how to weigh the measures used to achieve economic growth against the actual development realized, and we are left to rely on personal judgments concerning matters of social justice. Personally, I doubt that in the Brazilian case the ends justify the means, but this is a value judgment, not a scientific conclusion, and others will undoubtedly disagree.

C. *How to Determine Who Benefits from Economic Development*

Beyond these specific questions pertaining to the particular case of Brazil, this study raises a much more fundamental issue of general applicability, namely, *how should distributional concerns be brought to bear in evaluating a country's economic development?* In my 1975 paper, I showed that when a country's high income sector enlarges to absorb an increasing share of the population, the more rapid alleviation of poverty is invariably accompanied by greater measured inequality in the early stages. When growth takes place in this fashion, should rising inequality be interpreted as an economically meaningful "worsening" of the income distribution or as an emotively neutral statistical artifact inherent in the very nature of this class of relative inequality indices? I would tend to opt for the latter. In any case, the key point is that there is no necessary concurrence between absolute-income and relative-inequality based distributional studies.

What this all basically comes down to is whether we wish to give greater weight in our judgments about the distributional consequences of economic development to the alleviation of absolute poverty or to the narrowing of relative income inequality. Personally, I am most concerned about the alleviation of absolute poverty among the

[11] The desirability of a poverty focus has been stated clearly by Fishlow (1972, pp. 392–95) and Langoni (1972, pp. 80–81). For similar statements outside the Brazilian context, see, for instance, Robert McNamara in Hollis Chenery et al., and Dudley Seers.

very poorest, and have made use of measures with this explicit focus to the virtual exclusion of the rest of the income distribution. Others with different value judgments who may be more concerned than I with relative income comparisons or with the middle or upper end of the income distribution may wish to give relatively greater weight to changes in other measures in arriving at their own interpretations of the Brazilian experience.

Ideally, the choice of the measure to use and the specific welfare weights assigned should reflect the value judgments we wish to make. Despite recent advances in this area,[12] there is not yet any consensus on how best to go about bringing these judgments to bear in practice. This is perhaps one of the most important lessons emerging from the Brazil debate.

This reexamination of the Brazilian experience has raised some fundamental questions. Did the personal distribution of income *really* worsen in Brazil over the 1960's? Should the rising Gini coefficient weigh negatively in our welfare judgments, and if so, by how much? How much importance should be given to considerations of relative incomes as opposed to absolute poverty? By assigning heavy weight to changes in the usual indices of relative income inequality and interpreting these increases as offsets to the well-being brought about by growth, the participants in the Brazilian debate and others who have followed similar approaches in studies of other less developed countries appear to have overlooked important tendencies toward the alleviation of poverty.

[12]Among the recent works are Atkinson, Nicholas Stern, Montek Ahluwalia and Hollis Chenery, and the author and John Fei

REFERENCES

Irma Adelman and Cynthia T. Morris, *Economic Growth and Social Equity in Developing Countries*, Stanford 1973.

M. S. Ahluwalia and H. B. Chenery, "The Economic Framework," in Hollis B. Chenery et al., eds., *Redistribution with Growth*, New York 1974.

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

D. G. Champernowne, "A Comparison of Measures of Inequality of Income Distribution," *Econ. J.*, Dec. 1974, *84*, 787–816.

Hollis B. Chenery et al., *Redistribution with Growth*, New York 1974.

W. R. Cline, "Distribution and Development: A Survey of the Literature," *J. Develop. Econ.*, Feb. 1975, *1*, 359–400.

G. S. Fields, "On Inequality and Economic Development," disc. pap. 233, Econ. Growth Center, Yale Univ., Aug. 1975.

———, "More on Changing Income Distribution and Economic Development in Brazil," disc. pap. 244, Econ. Growth Center, Yale Univ. Apr. 1976.

——— **and J. C. H. Fei,** "On Inequality Comparisons," *Econometrica*, forthcoming.

A. Figueroa and R. Weisskoff, "Visión de las Pirámides Sociales: Distribución del Ingreso en América Latina," *Ensayos ECIEL*, Nov. 1974, *1*, 83–154.

A. Fishlow, "Brazilian Size Distribution of Income," *Amer. Econ. Rev. Proc.*, May 1972, *62*, 391–402.

———, (1973a) "Distribuição da Renda no Brazil: Um Novo Exame," *Dados*, 1973, *11*, 10–80.

———, (1973b) "Some Reflections on Post 1964 Brazilian Economic Policy," in Alfred Stepan, ed., *Authoritarian Brazil*, New Haven 1973, 69–113.

C. Furtado, "El Modelo Brasileño," *Trimestre Econômico*, July–Sept. 1973, *159*, 587–99.

C. Langoni, "Distribuição da Renda e Desenvolvimento Econômico do Brasil," *Estudos Econômicos*, Oct. 1972, *2*, 5–88.

———, "Income Distribution and Economic Development: The Brazilian Case," paper presented at World Econometric Society Congress, Toronto 1975.

———, "Review of Income Distribution Data: Brazil," disc. pap. 60, Res. Prog. in Econ. Develop., Woodrow Wilson School, Princeton Univ., Apr. 1975.

P. Malan and J. Wells, "Langoni e a Distribuição de Renda no Brasil," *Pesquisa e Planejamento Econômico,* Dec. 1973, *3,* 1103–24.

K. S. Mericle, "Corporatist Control of the Working Class: The Case of Post-1964 Authoritarian Brazil," in James M. Malloy, ed., *Authoritarianism and Corporatism in Latin America,* Pittsburgh 1976.

S. A. Morley and J. G. Williamson, "Demand, Distribution, and Employment: The Case of Brazil," *Econ. Develop. Cult. Change,* Oct. 1974, *23,* 33–60.

——, "Growth, Wage Policy and Inequality: Brazil During the Sixties," workshop pap. no. 7519, SSRI, Univ. Wisconsin, July 1975.

D. Seers, "The Meaning of Development," *Int. Develop. Rev.,* Dec. 1969, *11,* 2–6.

P. I. Singer, "Força de Trabalho e Emprego no Brasil, 1920–1969," *Centro Brasileiro de Análise e Planejamento,* São Paulo 1971.

N. H. Stern, "Welfare Weights and the Elasticity of the Marginal Valuation of Income," unpublished paper, St. Catherine's College, Oxford, Aug. 1973.

R. Weisskoff, "Income Distribution and Economic Wealth in Puerto Rico, Argentina, and Mexico," *Rev. Income Wealth,* Dec 1970, *16,* 303–32.

J. Wells, "Distribution of Earnings, Growth and the Structure of Demand in Brazil during the 1960's," *World Develop.,* Jan. 1974, *2,* 9–24.

J. P. Wogart, "Contrasting Employment Patterns in Brazil: (1940–1970)," mimeo, dept. of econ. and Inst. of Inter-American Stud., Univ. Miami, 1974.

Brasil, *Censo Demográfico: Resultados Preliminares,* Rio de Janeiro 1960.

——, *Tabulações Avançadas do Censo Demografia,* Rio de Janeiro 1970.

Brazilian Trends, São Paulo, April 1973.

Comisión Economica para América Latina, *Proyecto Sobre Medición y Análisis de La Distribución del Ingreso en Paises de América Latina, Tabulados de Trabajo, Brasil,* Documento Número E/CEPAL/L.115/8, Nov. 1974.

Fundação Getúlio Vargas, *Atualização Parcial do Sistema de Contas Nacionais, 1971–72,* June 1973, Rio De Janeiro.

U.S. Bureau of the Census, *Statistical Abstract of the United States, 1971,* Washington 1971.

World Bank, *The Assault on World Poverty,* Baltimore 1975.

Product Quality, Uncertainty, and Regulation: The Trucking Industry

By A. S. DE VANY AND T. R. SAVING*

The impact of uncertainty on optimal pricing and output decisions where quality is a parameter has been the subject of recent work with contributions by Robert Meyer, Gardner Brown and M. Bruce Johnson, and Phoebus Dhrymes. On the other hand, optimal pricing and product quality under certainty has been discussed extensively by Peter Swan and others. However, little if any work has been done on the determination of optimal pricing and quality under conditions of uncertainty. As we shall demonstrate below, the integration of the quality and price decision under uncertainty allows us to explain a broad class of economic behavior heretofore unexplained.

The central thesis of this paper is that industry capacity affects both the quantity and quality of output. Because of the impact of capacity on quality, pricing practices of certain industries, for example, value of service pricing in truck transport, become understandable as the result of normal competitive profit-maximizing behavior. Moreover, previous arguments that such pricing practices are not socially optimal become questionable.

The particular application of the quantity-quality relation considered here is the relation between capacity and waiting time in the truck transport industry.[1] This industry among others is faced with uncertainty on both the demand and supply sides. As a consequence of this uncertainty the system will be saturated at times and consumers must wait or forego use of the service. Thus, randomness gives rise to waiting time which functions as an implicit price. In this paper we construct a model of a competitive market that determines the equilibrium price, waiting time, capacity, and industry output.[2]

There are three central features in our model of this problem. First, waiting time fluctuates randomly, and there is imperfect information. This gives rise to search for short queues, and results in a process closely related to George Stigler's analysis of price search. Second, higher capacity reduces the probability that the system is overloaded, and thus reduces waiting time. As a consequence, there is economic value to excess capacity and the market must somehow trade off the probability of the system being empty (customers do not wait) against the probability of it being full (no wait for servers). This duality of the waiting time problem plays a central role in the determination of the optimal level of excess capacity.[3] Third, rational search leads to a stopping rule wherein searchers balk at joining queues longer than some reservation length. This process stabilizes queue lengths and ensures the existence of a steady-state equilibrium at the firm level. Industry equilibrium, however, requires

*Texas A&M University. This research was financed by the Motor Vehicle Manufacturers Association and by National Science Foundation Grant, SOC 76-06025. The views expressed here are not necessarily those of the MVMA. We wish to thank Armen Alchian and Hayden Boyd for comments and suggestions made on an earlier draft.

[1]Some work on the airline industry has considered this problem. See George Douglas and James Miller, and De Vany (1975a) for a discussion of the relation between schedule frequency and waiting time. Ross Eckert, Douglas, and De Vany (1975b) have considered industry supply and waiting time in the case of taxicabs. None of this work has considered the competitive provision of waiting time in a fairly general stochastic environment.

[2]In a sense the waiting time solution is determined in an implicit market. For an excellent discussion of implicit markets see Sherwin Rosen. For an example of waiting time and capacity utilization under monopoly, see De Vany (1976).

[3]This is not a new point, such duality has been forcefully argued by Armen Alchian among others. In addition to the transportation work cited, the point has been debated in the Soviet Union in the form of criticism of the full utilization doctrine. On this see Boris Gnedenko and Ivan Kovalenko.

that there be excess capacity in the aggregate. The expectation of the full price is a parameter at each firm, but the actual full price follows a random walk with a reflecting barrier below at the money price and a barrier above at the sum of this price and the maximum tolerable waiting cost.

I. The Traditional Analysis

The standard model of a competitive firm operating in multiple markets is that each firm maximizes a profit function of the form

$$(1) \qquad \pi = \Sigma p_i q_i - C(q_i; k)$$

where p_i and q_i are respectively the price and quantity in the ith market, and k is capacity. The relevant first-order conditions are

$$(2) \quad p_i + \sum_{j=1}^{n} p_j \frac{\partial q_j}{\partial q_i} - \left[C_{q_i} + \sum_{j=1} C_{q_j} \frac{\partial q_j}{\partial q_i} + C_k \left(\sum_j \frac{\partial k}{\partial q_j} \frac{\partial q_j}{\partial q_i}\right)\right] = 0; \quad i = 1, \ldots, n$$

In this model let one market be the dominant market, i.e., $q_d > q_i \forall i \neq d$. Assume that $(\partial k/\partial q_i) = 0$, $\forall i \neq d$, $(\partial q_j/\partial q_i) = 0$, and let $C_x = \partial C/\partial x$, $x = q_i, q_j, K$. Then (2) implies

$$(3) \qquad p_i = C_{q_i} \forall i \neq d$$

$$(4) \qquad p_d = C_{q_d} + C_k \frac{\partial k}{\partial q_d}$$

Thus, we get the traditional result that the nondominant markets pay only their direct marginal cost and the dominant market bears the full capital cost. That is, because the subsidiary or off-peak markets do not require any additional capital stock, these markets should not pay any of the capital cost.[4]

This result depends crucially on the assumptions that the markets are independent, and the marginal capital requirements for the off-peak markets are zero. In fact, as we shall show below the introduction of demand and supply uncertainty results in both of these assumptions being violated.

II. A Model of the Trucking Firm Under Uncertainty

Let the truck transport system consist of a road between points A and B and trucks capable of carrying unit loads. Let the shipping firms be expected profit maximizers so that mode of transport will be chosen on the basis of expected full price.[5] The full price of shipping a unit load is the actual transport charge between A and B plus the inventory holding cost associated with shipping time, where shipping time includes waiting time. The latter cost must be included because it represents a real cost of transport. If transport capacity increases, the volume of goods waiting for transit will fall and resources will be released. The expected cost of delivery time is then the product of the cost of holding a unit of inventories per unit of time and the expected total waiting time.[6]

Let the rate at which loads arrive for shipment at A and B be random variables whose distributions depend on the expected full price at A and B, respectively. Then

$$(5) \qquad q_A \sim f_A(q_A, P_A)$$

$$(6) \qquad q_B \sim f_B(q_B, P_B)$$

where q_A, q_B are the quantities of shipping demanded per period at A and B, P_A, P_B are the respective expected full prices, and f_A and f_B are probability density functions. The expected full prices are

$$(7) \qquad P_A = p_A + \eta_A \overline{W}_A$$

$$(8) \qquad P_B = p_B + \eta_B \overline{W}_B$$

[4]This traditional result is not affected by introducing price uncertainty since the expected profit maximization simply implies the substitution of expected price for certain prices in (1).

[5]The theory of the shipper's demand for truck transportation is developed in the authors (1975).

[6]Thus goods in transit cannot be used by anyone and represent a loss to the system. This is not deadweight loss, however, since to reduce goods in transit in an optimal situation would result in an increase in shipping resources by more than the saving in inventory holding costs.

where p_A and p_B are the actual transport charges per load at A and B, η_A and η_B are the costs of holding a unit (one truck load) of inventory at A and B, $\overline{W}_A$ and $\overline{W}_B$ are the expected elapsed times from transport order to delivery. The full prices equal transport price plus the alternative cost of inventory necessitated by the shipping process. For simplicity let us treat each commodity as a separate competitive market so that at point A, η_A is equal for all shippers, and similarly at point B, η_B is equal for all shippers.[7]

A. *Firm Demand*

Since we are treating each route as a separate competitive market, each firm is a full-price taker at both A and B.[8] So long as each firm maintains the market expected full price, it behaves as if the arrival rate at its door is proportional to its capacity. That is, a doubling of capacity will double the expected number of shipment arrivals. Essentially, this assumption is a stochastic version of the usual competitive assumption that a firm can sell any desired output at the market equilibrium price. In this case firms can have any desired expected number of customer arrivals by simply supplying the appropriate capacity.[9]

While the expected full price is identical across firms before a customer begins search, once he arrives at a particular firm the information obtained yields a conditional expected full price. In particular, since arrivals are random the queue length will differ across firms and time, implying that the conditional expected full price will not equal the expected full price. This difference in conditional expected full price across firms opens up the possibility that individual shippers at an instant in time will find the queue length at a particular trucking firm too long, i.e., the conditional expected price is too high, and they search for a shorter queue.[10]

Let the conditional expected full price $(P \mid Q)$ be

$$(P \mid Q) = p + \eta(W \mid Q) \tag{9}$$

where $(W \mid Q)$ is the expected wait given that the queue at a particular time is of length Q, and we have suppressed the market subscripts since the discussion applies to both A and B. The expected cost of not shipping with the first firm is the market expected full price plus the marginal cost of search σ. Thus, the customer will balk at the queue any time

$$\sigma < \eta[(W \mid Q) - \overline{W}] \tag{10}$$

i.e., when search costs are less than the value of the difference between the conditional and unconditional expected waits. The maximum queue that a customer will tolerate is Q such that (10) is an equality. This maximum queue results in a maximum or reservation expected full price R which is

$$R = p + \eta\overline{W} + \sigma \tag{11}$$

Given the reservation expected full prices (R_A, R_B) at A and B and the distribution of conditional expected full prices, there will exist a mean number of searches per customer m_A, m_B at each termination point A and B.[11] Given that the level of this

[7] We assume that the inventory costs are paid by the shipper. Actually, the burden of the inventory costs is shared by shippers and receivers of goods. In our 1975 paper we also show the shipper's capacity cost enters into his demand for shipping capacity.

[8] If a single firm operates in several markets, we assume that the firm's total capacity is equal to the sum of its individual market capacities so that there are no benefits from risk pooling. Such risk pooling is one reason for mergers by firms with different routes. The extent of these economies is an important issue in understanding the rate of merger activity. This is, of course, not the problem being discussed here.

[9] In effect we are assuming that customers search over trucks on a random basis. Thus, arrivals at any one firm will be proportional to that firm's capacity.

[10] This behavior is called balking in the queueing theory literature (see Frank Haight).

[11] The relation between the optimal acceptance price and the mean number of searches is given by Lester Telser. Here we assume that the acceptance full price lies far enough below the (infinite) upper bound on the distribution of full prices generated by fluctuating queue lengths so as to generate positive net returns to search, i.e., $m_A, m_B > 1$. If $m_A, m_B = 1$, what Telser calls the naive rule will hold and no balking will occur.

search activity is independent of total arrivals and that the arrival rate of customers at a single firm is proportional to that firm's capacity, it follows that the expected number of customer arrivals for the ith firm at A and B, a^i, b^i, respectively, are

$$a^i = \frac{m_A N_i}{N} a \tag{12}$$

$$b^i = \frac{m_B N_i}{N} b \tag{13}$$

where a and b represent industry expected arrivals at A and B, respectively, and N is total industry capacity.[12]

Now assume that the arrival of searchers to the industry is Poisson so that each firm's arrivals will be Poisson distributed and in particular[13]

$$q_A^i \sim \frac{(a^i t)^q}{q!} e^{-a^i t} \tag{14}$$

$$q_B^i \sim \frac{(b^i t)^q}{q!} e^{-b^i t} \tag{15}$$

B. *Firm Supply*

On the supply side we assume that the trucking firms of size N_i locate at the point with the greatest traffic flow.[14] Without loss of generality let this point be A. Assume that each firm sends trucks to B as loads arrive for shipment and that these trucks unload at B, load if a shipment is waiting, and return to A.[15] Thus, we assume that empty trucks move from B to A but that no empties move from A to B.[16] Denote the number of trips that can be completed per period as n, and let n be random and distributed

$$n \sim h(n, N_i) \tag{16}$$

with expected value μN_i.[17] Thus, the expected time for a truck to load at A, make the trip to and unload at B, load if a shipment is waiting and return to A, is $(1/\mu)$. Now assume that n in (16) is Poisson so that

$$n \sim \frac{(N_i \mu t)^n}{n!} e^{-N_i \mu t} \tag{17}$$

Thus, the expected number of trips is μN_i as assumed in (16). In addition, assume that the number of round trips actually completed and the number that can be completed are equal.[18]

[12]As we shall point out below, the flow of traffic at B will not necessarily be proportionate to (N_i/N). Thus, the capacity of the firm at B may not equal N_i. This assumption, however, considerably simplifies the subsequent analysis.

[13]What we require is (i) the probability of an industry arrival in an interval $(t, t + \Delta t)$ is equal to $\alpha \Delta t$ plus a term that goes to zero faster than Δt, (ii) the probability of more than one arrival in the interval $(t, t + \Delta t)$ goes to zero faster than Δt; and (iii) arrivals in nonoverlapping intervals are independent. These assumptions will result in the distribution of demands being Poisson and allow us to utilize the many convenient results of queueing theory. More general results may follow from a more general distribution of arrivals and service rates.

[14]This assumption implies that there are no economies of scale in expanding routes or commodities. While we are taking some liberty with the facts, this assumption greatly simplifies the analysis and as we shall demonstrate below, we give up very little in generality.

[15]Here we want to consider situations where one market dominates the other, a common element in actual trucking. This will turn out to be a very fruitful way of analyzing certain regulatory actions.

[16]Note we are ignoring the less than truckload shipment, or we are assuming that there is an intermediate step, freight forwarding if you like, where all loads are consolidated into unit truckloads. The model can be extended to allow a stock of trucks to be held at both points and for empty trucks to move in either direction. This addition adds considerably to the complexity, however, with no gain for our purposes

[17]Essentially we are assuming that the number of trips per truck are independently distributed with mean μ. Then $N_i\mu$ is the mean of the sum of N_i independent random variables each with mean μ. In the queueing literature $N_i\mu$ is referred to as the service rate.

[18]Essentially we are approximating the N-server queueing model, each server having capacity μ ($(M/M/N)$ in David Kendall's notation) with a single-server model with capacity $N\mu$. The expression for expected wait for the N-server model is

$$\overline{W} = \pi \int_0^\infty t(N\mu - \alpha) e^{-(N\mu - \alpha)t} dt$$

where π = probability all N servers busy. On this see Gnedenko and Kovalenko. Evaluate to get $\overline{W} = \pi/N\mu - \alpha = \pi/N\mu(1 - \rho)$ where $\rho = \alpha/N\mu$. Thus the single-server model approximates the N-server model expected wait whenever ρ is close to π. This approximation is suggested by the fact that the fraction of time during which a single server of capacity $N\mu$ is busy (ρ) is a

III. The Steady-State Solution

As we discussed above, shippers search across firms until they find one with an acceptable conditional expected full price, i.e., a firm where the queue is sufficiently short. The shipper then chooses the first firm encountered with a conditional expected full price less than or equal to his reservation full price. From the definition of the full price, this implies that the maximum expected waiting times which a searcher will tolerate at firm i, ω_{Ai} and ω_{Bi}, are

$$(18)\qquad \omega_{Ai} = \frac{R_A - p_A^i}{\eta_A} = \overline{W}_A + \frac{\sigma}{\eta_A}$$

$$(19)\qquad \omega_{Bi} = \frac{R_B - p_B^i}{\eta_B} = \overline{W}_B + \frac{\sigma}{\eta_B}$$

where p_A^i, p_B^i are the ith firm's transport charges at A and B.

A. *The Solution at A*

Let the potential customers of firm i at point A estimate the total shipping time from observation of the queue length at the time of search and from knowledge of μN_i. Given that there are Q_A loads at firm i the conditional expected wait is

$$(20)\qquad (W_A \mid Q_A) = \frac{Q_A}{N_i \mu}$$

That is, the conditional expected wait is the number at firm i, Q_A, times the expected service time $(1/N_i\mu)$. Thus, the maximum number at firm i that a searcher will tolerate is $\Gamma_{Ai} = \omega_{Ai} N_i \mu$. The shipper then stops at the first firm encountered who has $(\Gamma_{Ai} - 1)$ or fewer in its system.[19] Given the arrival distribution (14) and the service distribution (17) it can be shown that the expected number in the system at firm i in the steady state is

$$(21)\qquad \overline{Q}_{Ai} = \frac{r_{Ai}}{1 - r_{Ai}} - (\Gamma_{Ai} + 1)\frac{r_{Ai}^{\Gamma_{Ai}+1}}{1 - r_{Ai}^{\Gamma_{Ai}+1}}$$

where $r_{Ai} = a^i/N_i\mu$.[20] The necessary and sufficient condition for the existence of the above steady-state solution is that the probability that the system is empty be positive.[21]

The probability that a shipper will balk at firm i is equal to the probability that there are Γ_{Ai} loads in the system, which is

$$(22)\qquad \gamma_{Ai} = \frac{(1 - r_{Ai}) r_{Ai}^{\Gamma_{Ai}}}{1 - r_{Ai}^{\Gamma_{Ai}+1}}$$

where γ_{Ai} is the probability of a balk. Given that the fraction γ_{Ai} of arrivals balk, realized shipments per unit of time are

$$(23)\qquad \alpha^i = a^i(1 - \gamma_{Ai})$$

and the effective utilization rate of the firm's capacity is

$$(24)\qquad \rho_{Ai} = r_{Ai}(1 - \gamma_{Ai})$$

The steady-state probability of an empty system is $(1 - \rho_{Ai})$ which must be positive for existence. Thus, it follows that $\rho_{Ai} < 1$. In other words, the mean effective number of shipments must be less than the firm's mean capacity if steady-state equilibrium exists. Thus, we have a competitive capacity utilization theorem which says the firm cannot on average fully utilize capacity.[22]

B. *The Solution at B*

The steady-state solution for the system at point B is constructed in a fashion similar

good approximation to the probability that the proportion of N servers, each with capacity μ, is busy. For more general arrival and service distributions there is little distinction between N servers and one server, once again see Gnedenko and Kovalenko.

[19]When a searcher begins search, all trucking firms have identical expected full prices. Upon arrival at a particular firm the searcher has information not available before search; the number in the queue. Thus with the additional information he decides whether to join the queue or search further.

[20]We can obtain the solution to this problem from Haight's model or by recognizing that if everyone balks when there are Γ_{Ai} loads in the system, then this is equivalent to the Markovian single-server model $(M/M/1/\Gamma_{Ai})$ where Γ_{Ai} is some absolute limit on the number that can be in the system (see Donald Gross and Carl Harris).

[21]This condition will be assured if Γ_{Ai} is finite. See Haight or De Vany (1976).

[22]To attempt to do so would be inefficient since it would impose large waiting time costs on shippers.

to the solution at point A. Here we assume that trucks unload at B and pick up a load if one is available. If no load is available, the truck returns to A empty.[23] From these assumptions and the steady-state solution for the firm at A, it follows that the arrival rate of trucks at B is α^i.[24] Assuming trucks leaving A are Poisson, then arrivals at B are Poisson with arrival rate α^i so that

$$(25)\qquad n_B \sim \frac{(\alpha^i t)^n}{n!} e^{-\alpha^i t}$$

From the fact that the service rate at B is α^i and (19) it follows that the maximum number of shipments in firm i's system which will be tolerated by a shipper at B is

$$(26)\qquad \Gamma_{B_i} = \alpha^i \omega_{B_i}$$

Using (15) and (25) the steady-state solution for the expected number in firm i's system at B is

$$(27)\quad \bar{Q}_{B_i} = \frac{r_{B_i}}{1 - r_{B_i}} - (\Gamma_{B_i} + 1)\frac{r_{B_i}^{\Gamma_{B_i}+1}}{1 - r_{B_i}^{\Gamma_{B_i}+1}}$$

where $r_{B_i} = b^i/\alpha^i$. As above, the probability that a shipper at B will balk at firm i is the probability that the queue length is Γ_{B_i} which is

$$(28)\qquad \gamma_{B_i} = \frac{(1 - r_{B_i})r_{B_i}^{\Gamma_{B_i}+1}}{1 - r_{B_i}^{\Gamma_{B_i}+1}}$$

where γ_{B_i} is the probability that a shipper balks. Given (15) the expected realized flow of shipments from B to A for firm i is

$$(29)\qquad \beta^i = b^i(1 - \gamma_{B_i})$$

and the effective load on firm i relative to B is

$$(30)\qquad \rho_{B_i} = r_{B_i}(1 - \gamma_{B_i}) = \frac{\beta^i}{\alpha^i}$$

C. *Some Additional Comments*

Normally one would expect that stability of a queueing process would require that the arrival rate be less than the ability of the system to service the arrivals. In the case where the shippers balk, however, the number of searchers may exceed the capacity of the firm. The reason for this is that arrivals balk when the queue reaches the critical value, and therefore the length of the line stabilizes and becomes independent of the history of the system. At the industry level, however, we have assumed that balking does not occur and so industry stability requires that the expected industry arrival rate a be less than industry capacity $N\mu$, where $N = \Sigma N_i$.

The proportion of searchers who balk depends on the capacity of the firm relative to the search rate. The search rate is determined by the full price and the distribution of waiting times, as well as the cost of search. Since the purchase of shipping services is a repetitive process with low search cost, we assume that search drives expected full price to equality among firms Even in this case, however, there is random variation in the full price since it depends on the number in the queue at the time of search. So, with equal expected full prices across the market, there will be search for short queues.

IV. Equilibrium Conditions for the Firm

Before we can derive the equilibrium conditions for the firm, we must derive the relation between the expected net flow of shipments α^i, β^i, and the choice variables of the firm. Since demand depends only on expected full price, any firm whose expected full price is above the expected full price of other firms will receive no business. Competition then implies that the firms are expected full-price takers. A competitive firm is not able to vary its price and output rate freely, but is subject to the constraint that its full price remain equal to the market-clearing full price. This means that if a firm changes its price, expected wait must change in an offsetting manner so as to leave full price unaffected. If capacity is changed at constant output, then the resulting effect on expected wait must be offset by a change in price. Essentially what occurs is

[23]Additionally, we are assuming that loading and unloading time is included in the service time per truck μ.

[24]This is a well-known property of serial queues; see Gross and Harris.

that the net arrival rate at the firm adjusts so that no action of the firm can result in its expected full price deviating from the market determined expected full price.

From (22), (23), and (28), (29), we can write α^i and β^i in general form resembling a demand function as

$$\alpha^i = \alpha^i(\overset{-}{p^i_A}, \overset{+}{N_i}) \tag{31}$$

$$\beta^i = \beta^i(\overset{-}{p^i_B}, \overset{+}{\alpha^i}) \tag{32}$$

where (31) and (32) are derived holding full price constant and the signs over the arguments are the signs of the partial derivative with respect to that argument. The noted signs are easily demonstrated by first recognizing that since full price is a parameter for the firm, increases in the service rates (N_i in (31) and α^i in (32)) for fixed p^i_A and p^i_B must result in increased utilization rates so that expected waiting time remains unchanged. Secondly, increases in transport prices (p^i_A, p^i_B) must result in decreases in utilization rates in order that expected waiting time falls enough to maintain full price. From the definition of full price it follows that these derivatives may be written as

$$\frac{\partial \alpha^i}{\partial p_A} = -\frac{1}{\eta_A \dfrac{\partial \overline{W}_A}{\partial \alpha^i}} < 0 \tag{33}$$

$$\frac{\partial \alpha^i}{\partial N_i} = -\frac{\dfrac{\partial \overline{W}_A}{\partial N_i}}{\dfrac{\partial \overline{W}_A}{\partial \alpha^i}} > 0$$

$$\frac{\partial \beta^i}{\partial p_B} = -\frac{1}{\eta_B \dfrac{\partial \overline{W}_B}{\partial \beta^i}} < 0 \tag{34}$$

$$\frac{\partial \beta^i}{\partial \alpha^i} = -\frac{\dfrac{\partial \overline{W}_B}{\partial \alpha^i}}{\dfrac{\partial \overline{W}_B}{\partial \beta^i}} > 0$$

Now using (31) and (32) write the profit function for the firm as

$$\pi_i = p^i_A \alpha^i(p^i_A, N_i) + p^i_B \beta^i(p^i_B, \alpha^i) - C(\alpha^i, \beta^i, N_i) \tag{35}$$

where $C(\cdot)$ is the expected cost function with $\partial C/\partial \alpha^i = C_{\alpha^i} > 0$, $\partial C/\partial \beta^i = C_{\beta^i} > 0$ and $\partial C/\partial N_i = C_{N_i} > 0$.[25,26] As we have pointed out above competition implies that the firms are full-price takers. From the definition of full price and the solutions for expected wait it follows that the firm can choose either its quantities (α^i, β^i) and its capacity N_i or its prices (p^i_A, p^i_B) and its capacity N_i, but not both. That is, once the quantities or the prices are chosen, the market determined full prices determine the remaining variables.

Even though we are assuming competition, it is not the case that the firms are transport price takers. Indeed a firm can have any set of transport prices it chooses but must accept the resulting market impact on its output since its waiting time must adjust until its full price is at the market level. It is perfectly legitimate then, for us to let p^i_A, p^i_B, and N_i be the choice variables for the firm, and assuming π_i is concave in these variables, the first-order conditions which characterize an optimum are

$$\left(p^i_A \frac{\partial \alpha^i}{\partial p^i_A} + \alpha^i\right) + p^i_B \left(\frac{\partial \beta^i}{\partial \alpha^i}\right)\left(\frac{\partial \alpha^i}{\partial p^i_A}\right) - \left(C_{\alpha^i} + C_{\beta^i} \frac{\partial \beta^i}{\partial \alpha^i}\right) \frac{\partial \alpha^i}{\partial p^i_A} = 0 \tag{36}$$

$$\left(p^i_B \frac{\partial \beta^i}{\partial p^i_B} + \beta^i\right) - C_{\beta^i} \frac{\partial \beta^i}{\partial p^i_B} = 0 \tag{37}$$

[25] The function $C(\cdot)$ is the expected cost function. Since the Poisson is a one-parameter distribution, expected costs will depend only on α^i, β^i, and N_i. With more general arrival and service distributions the expected cost function would also be a function of the other parameters of the relevant probability density functions unless the cost function is linear.

[26] It may appear that C_{N_i} is necessarily zero since it must vanish if chosen so that costs are a minimum for any given vector of outputs. This would be the case if product quality was independent of capacity. In our case, however, an increase in capacity for given levels of output increases quality by reducing waiting time. Accordingly, at the market determined full prices the transport prices at the firm are positively related to capacity, i.e., capacity has a positive marginal revenue. Thus, the firm will in equilibrium not choose that level of capital which minimizes costs for given output. Instead it must weigh the cost effect of additional capital against the revenue effects and will choose a level of capital such that $C_{N_i} > 0$.

$$(38)\quad \left(p_A^i + p_B^i \frac{\partial \beta^i}{\partial \alpha^i}\right)\frac{\partial \alpha^i}{\partial N_i} - \left(\left(C_{\alpha^i} + C_{\beta^i}\frac{\partial \beta^i}{\partial \alpha^i}\right)\frac{\partial \alpha^i}{\partial N_i} + C_{N_i}\right) = 0$$

By solving (36)-(38) for p_A^i, p_B^i, and C_{N_i} and using (33), (34), we can write the necessary conditions as[27]

$$(39)\quad p_A^i = C_{\alpha^i} + \alpha^i \eta_A \frac{\partial \overline{W}_A}{\partial \alpha^i} - \beta^i \eta_B \frac{\partial \overline{W}_B}{\partial \alpha^i}$$

$$(40)\quad p_B^i = C_{\beta^i} + \beta^i \eta^B \frac{\partial \overline{W}_B}{\partial \beta^i}$$

$$(41)\quad C_{N_i} = -\alpha^i \eta_A \frac{\partial \overline{W}_A}{\partial N_i}$$

Since the p_B^i result is simplest, let us begin with market B. Note that the optimal price in B is greater than the marginal cost of output at B. In fact, the second term on the right-hand side of (40) measures the rate of decrease in the price at B per unit increase in the net traffic flow at B necessary to compensate for the increased waiting time associated with an increase in the net traffic flow. Thus, the price in B reflects a charge for the marginal cost of output C_{β^i} and a congestion toll equal to the value at B of the increase in waiting time associated with one additional unit of output.[28]

In a similar fashion the terms $[-\beta^i \eta_B(\partial \overline{W}_B/\partial \alpha^i)]$ and $[\alpha^i \eta_A(\partial \overline{W}_A/\partial \alpha^i)]$ are the respective values of the decreased waiting time at B and increased waiting time at A that result from a unit increase in the net traffic flow at A.[29] The price at A, then, is equal to marginal cost at A plus a congestion toll at A less the value of the reduced congestion at B. Thus, the greater the benefits of additional capacity at B, the lower the price at A.

Both prices p_A and p_B exceed the direct marginal cost of transportation. Why then is it not in the interest of the firm to attempt to increase output at A and B? The answer to this question is that in order to increase output the firm must add to its capital stock or reduce its transport price. In fact, the prices at A and B just equal the full marginal cost (direct marginal cost plus the required charges in transport prices) of an additional unit of output. Alternatively, prices at A and B equal their respective direct marginal cost plus the additional capital stock required to maintain transport prices.

Equation (41) indicates that even though the price at B includes a congestion toll charge, the stock of capital is determined exclusively in the market at A. We should point out, however, that the choice of p_A^i and p_B^i both impact on the net traffic flow at A. Since both p_A^i and p_B^i are influenced by traffic at B, this traffic indirectly affects the choice of N_i. The optimal prices at A and B are then fixed to adjust net traffic flows so that each purchaser of transport pays a congestion toll equal to the value of the increased waiting time associated with his use of the system and accordingly pays a share of the capital cost.

The conditions (39)-(41) plus the zero-profit conditions imply that the competitive outcome is efficient in the context of no highway congestion which we have assumed. Price is equal to marginal social cost in market A and market B and the weighted marginal value of output equals its weighted marginal cost with marginal cost also equaling the foregone value of the resources employed by the trucking industry, which are measured by average cost. This conclusion, of course, requires modification

[27] The resulting equations are not reduced form equations for p_A^i and p_B^i, since terms on the right are functions of price. The equations, however, do indicate the relations among price and the other variables in equilibrium.

[28] We are using congestion toll as a synonym for the marginal time cost. These costs are the equivalent of the marginal effect on highway congestion of additional users. Hence, the term congestion toll seems appropriate. Note, also, the arrival distribution is stationary so the toll is not the result of a periodic demand distribution. The peak is randomly distributed over time intervals, rather than fixed.

[29] Thus if waiting time is considered the variable input, we know that total costs are minimized since the cost of an additional unit of capacity services equals the value of the variable inputs saved if output is produced with another unit of capacity (see Herbert Mohring).

if there is highway congestion coupled with inefficient pricing of highway use.

The above results have important implications outside the transport industry. For example, markets A and B are analogous to peak and off-peak period markets.[30] Even though capacity is scaled primarily to market A, the firm charges a congestion toll in market B as well as in market A. The reason for this lies in the fact that effective demand for the service even in the off-peak market depends on the rate of utilization of capacity in that period. If one additional shipper joins the system at B he increases the expected wait for all other users because ρ_{B_i} increases. In the standard analysis of peak load pricing, capacity in the off-peak period has no effect on the effective demand for the service. It therefore follows from the standard analysis that only peak capacity should be rationed through congestion pricing. In the setting of substantial inventory holding cost and uncertainty that we have been dealing with, the standard model of the peak load pricing problem clearly does not apply. In fact, it would seem that the standard analysis is inapplicable to a wide class of service industries where effective demand depends on waiting time, or any related stochastic quantity such as reliability.[31]

V. Industry Equilibrium

Without loss of generality we can assume that all firms are identical. Then we can use (39)–(41) to solve for $\hat{\alpha}, \hat{\beta}, \hat{N}$, the common values of each firm's optimum α^i, β^i, and N_i. Each of these variables will be a function of the full prices, so we write

$$\hat{\alpha} = \hat{\alpha}(P_A, P_B) \tag{42}$$

$$\hat{\beta} = \hat{\beta}(P_A, P_B) \tag{43}$$

$$\hat{N} = \hat{N}(P_A, P_B) \tag{44}$$

From Section II, (5) and (6), we have that the expected arrivals at the industry level a and b are functions of the full prices at A and B, respectively, so that

$$a = a(P_A) \tag{45}$$

$$b = b(P_B) \tag{46}$$

Additionally, from (7) and (8) we have

$$P_A = p_A + \eta_A \overline{W}_A \tag{47}$$

$$P_B = p_B + \eta_B \overline{W}_B \tag{48}$$

Total industry supply is simply

$$\alpha^s = f\hat{\alpha}(P_A, P_B) \tag{49}$$

$$\beta^s = f\hat{\beta}(P_A, P_B) \tag{50}$$

where f is the number of firms. In addition, industry equilibrium requires that profits be zero so that

$$p_A\hat{\alpha} + p_B\hat{\beta} - C(\hat{\alpha}, \hat{\beta}, \hat{N}) = 0 \tag{51}$$

Given that all firms are identical, the expected waiting time at the industry must equal the common expected waiting time at each firm. For each pair (P_A, P_B) of full prices there is a pair $(\overline{W}_A, \overline{W}_B)$ of supplied expected waiting times derivable from the solution of (20)–(21), (34), and (35). Thus, we can write

$$\overline{W}_A = \overline{W}_A(P_A, P_B) \tag{52}$$

$$\overline{W}_B = \overline{W}_B(P_A, P_B) \tag{53}$$

Market equilibrium requires

$$\alpha(P_A) = f\hat{\alpha}(P_A, P_B) \tag{54}$$

$$\beta(P_B) = f\hat{\beta}(P_A, P_B) \tag{55}$$

Additionally define total industry capacity as

$$N = f\hat{N} \tag{56}$$

From (54) and (55) we can solve for P_A and P_B, which in turn determine all the other variables in the system. Thus, we have a model which determines the capacity of the industry and the transport costs. Given the demand equations and the service time, the expected wait is then determined and also the full price. Given the full price, we have the net flow of traffic which in con-

[30] The analogy is not an exact one since there is no backhaul across peak and off-peak markets, but there is some cross-price elasticity of demand. In competitive markets we will have price-taking behavior in the peak and off-peak markets at the firm level, with the cross-substitution effect occurring at the market level. We are working on a solution to this problem.

[31] For an examination of peak load pricing in a context where reliability is jointly determined with effective demand, see the authors (1976).

junction with the previously determined industry capacity gives us the utilization rate, or the equilibrium amount of "excess" capacity.

VI. Competitive Pricing: Some Implications

In this section we want to summarize the view of competitive pricing which the model affords and to show the applicability of this view to some problems in truck transport rate setting. We have shown that competition results in equilibrium prices that contain a congestion toll in addition to a fee covering the marginal cost of output. Each customer pays the full expected social marginal cost of output. In effect the firm does not allow its capacity to be treated as a free access resource by customers and instead levies a crowding charge on every user of his capacity equal to the marginal expected waiting cost. This waiting cost is directly related to the time value of customers.

It is well known that trucking firms charge rates which differ among commodity classes according to the value of the shipment. The usual explanation of this phenomenon is that since demand elasticity is inversely related to the value of product, the trucking industry is practicing discriminatory pricing. This conclusion rests on the erroneous assumption that competition would price only on the basis of marginal hauling cost and would not differentiate rates on the basis of the value of shipment. Prior analyses of the trucking industry have neglected the competitive provision of service quality entirely and so have failed to discover the correct pricing formulas of the competitive process.

By examining equations (39)-(41) we can see that a competitive trucking industry prices partly on the basis of the value of the shipment. Prices at both A and B include the marginal hauling cost plus a charge reflecting the value of the marginal waiting time. This latter charge results from the fact that increased waiting time increases inventory holding costs.

From the above it is clear that relative prices which differ from relative marginal hauling costs are not necessarily discriminatory. The issue is whether or not actual pricing behavior in the trucking industry results in relative prices which are different from those that would have resulted from a competitive environment. The existing empirical work has failed to come to grips with this issue since it assumes that relative prices which differ from relative marginal hauling costs imply discriminatory pricing behavior.[32]

There is widespread belief that rates should be related to distance of haul in some reasonably uniform manner. The argument states that the rate per mile should make allowance for fixed costs and then decline with distance since marginal hauling costs per mile decline with length of haul.[33] Our model implies that a rate structure of this form is inefficient and will penalize long-haul markets with poor service. It can also be shown that the logic of the rate formula is defective since it offers no basis for determining the optimal level of fixed capacity on which to base the intercept of the rate equation.

We can shed some light on this problem by considering a competitive industry without balking.[34] Under these circumstances the competitive waiting time is

$$\overline{W}_A = \frac{1}{\hat{N}\mu - \hat{\alpha}} \tag{57}$$

$$\overline{W}_B = \frac{1}{\hat{\alpha} - \hat{\beta}} \tag{58}$$

The appropriate marginal waiting times are then

$$\frac{\partial \overline{W}_A}{\partial \hat{\alpha}} = \frac{1}{(\hat{N}\mu - \hat{\alpha})^2} \tag{59}$$

[32] In fact, Josephine Olson uses evidence that relative prices differ from relative marginal hauling costs to conclude that the trucking industry is not competitive.

[33] The testimony of Richard Hinchcliff indicates this is precisely how rates are set. This uniform rate structure is closely tied to the notion of nondiscriminatory rates.

[34] Similar results can be shown for the balking case, but only after considerable algebraic manipulation.

$$\frac{\partial \overline{W}_B}{\partial \hat{\beta}} = \frac{1}{(\hat{\alpha} - \hat{\beta})^2} \tag{60}$$

which are both positive.

Now changes in trip distance change μ so that the relevant derivatives are

$$\frac{\partial^2 \overline{W}_A}{\partial \hat{\alpha} \partial \mu} = - \frac{2N}{(N\mu - \alpha)^3} \tag{61}$$

$$\frac{\partial^2 \overline{W}_B}{\partial \hat{\beta} \partial \mu} = 0 \tag{62}$$

Since we have assumed that the arrival rate per firm at A is unaffected by distance, the marginal waiting time at B remains unchanged. But marginal waiting time at A is negatively related to μ. In other words, marginal waiting time increases with distance. Thus, additional capacity reduces waiting time by more on long routes than on short routes. Given two routes of equal density, the route of greater distance will operate with more trucks, lower utilization rates, and higher prices provided the cost of hauling is the same.

The cost of hauling over routes of differing distance will, of course, usually differ. The above proposition illustrates only the effect of the demand side on capacity. If they are shipping equal-value loads, shippers on long-haul routes will have higher inventory and production costs than shippers operating over short-haul routes. Long-haul shippers will, therefore, pay a higher premium for capacity in the trucking industry than will short-haul shippers, and the premium is higher the longer the shipping distance. If the capacity premium rises more rapidly with distance than does the marginal hauling cost, then long-haul routes will operate with a larger fleet of trucks relative to number of loads shipped than will short-haul routes. Average cost per load mile will not decline with distance if the shipper's premium for capacity is such as to bid down the truck utilization rate sufficiently to offset the low hauling cost per mile. The ratio of waiting time to total shipping time will be lower the longer the route and the ratio of price to marginal cost higher. Thus, if rates are based on average or marginal hauling cost per load mile, then longer routes may receive poorer service than they would receive under competitive pricing.

VII. Concluding Remarks

While the competitive firm in the usual constant-quality certainty theory is assumed to be a product-price taker, this assumption must be modified when uncertainty and variable quality are present. What we have shown here is that the appropriate assumption for the uncertain variable quality case is to require the firm to be a full-price taker. This full price as a parameter model implies that the firms are waiting-time takers for any level of transport price. That is, if the firm takes any action which, given its transport price, changes its waiting time, the net arrival rate changes until the market determined level of waiting time is restored.

Within the context of this approach we have derived the equilibrium conditions for a competitive firm operating on a road between two points. The resulting optimal pricing structure implies that shippers at the higher traffic point pay the marginal hauling cost plus the marginal congestion cost at that point less the reduction in congestion at the low traffic point. The magnitude of these charges is directly proportional to the value of time at the two points.

Significantly, the shippers at the lower traffic point still pay a congestion charge in addition to the marginal hauling cost. That is, even here an additional shipper imposes a cost on the other users of the system by increasing the probability that they will have to wait and by increasing the expected length of wait when they do wait. These results are in direct contradiction to the usually asserted optimal pricing for such a system wherein it is claimed that the shipper at the low traffic point should pay only marginal hauling cost.

The reason our results differ so from the traditional results is the fact that we have explicitly accounted for the effect of the

mean flow through the system on quality. In the case considered here this quality was represented by waiting time. Thus an increase in the rate of use of the system decreases its quality for fixed capacity since it increases the expected wait. Moreover, this deterioration in quality occurs even when traffic intensity rises at the low traffic point.

The fact that the congestion charge is positively related to the value of time indicates that even if the marginal hauling costs of two commodities are equal, their respective equilibrium transport charges will in general be unequal. In fact, the commodity with the higher value per unit load will pay higher transport charges and receive in turn lower waiting time or, put differently, a higher quality product. In this way a competitive truck transport system minimizes the total cost of transporting the sum of all commodities.

REFERENCES

A. A. Alchian, "Information Costs, Pricing, and Resource Unemployment," *Western Econ. J.*, June 1969, *7*, 109–28.

G. Brown and M. B. Johnson, "Public Utility Pricing and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119–28.

A. S. De Vany, (1975a) "The Effect of Price and Entry Regulation on Airline Output, Capacity and Efficiency," *Bell J. Econ.*, Spring 1975, *6*, 327–45.

———, (1975b) "Capacity Utilization Under Alternative Regulatory Restraints: An Analysis of Taxi Markets," *J. Polit. Econ.*, Feb. 1975, *83*, 83–95.

———, "Uncertainty, Waiting Time and Capacity Utilization—A Stochastic Theory of Product Quality," *J. Polit. Econ.*, June 1976, *84*, 523–41.

——— **and T. R. Saving,** "Truck Transportation Efficiency," report for the Motor Vehicle Manufacturers Assn., Summer 1975.

——— **and** ———, "Uncertainty, Reliability and Peak-Load Pricing," unpublished paper, Texas A&M Univ. 1976.

P. Dhrymes, "On the Theory of the Monopolistic Multi-Purpose Firm Under Uncertainty," *Int. Econ. Rev.*, Sept. 1964, *5*, 239–57.

George W. Douglas, "Price Regulation and Optimal Service Standards: The Taxicab Industry," *J. Transp. Econ. Policy*, May 1972, *6*, 116–27.

——— **and James C. Miller III,** *Economic Regulation of Domestic Air Transport: Theory and Policy*, Washington 1973.

R. D. Eckert, "On the Incentives of Regulators: The Case of Taxicabs," *Publ. Choice*, Spring 1973, *14*, 83–101.

Boris V. Gnedenko and Ivan N. Kovalenko, *Introduction to Queueing Theory*, Jerusalem 1968.

Donald Gross and Carl M. Harris, *Fundamentals of Queueing Theory*, New York 1974.

F. M. Haight, "Queueing with Balking," *Biometrika*, Dec. 1957, *44*, 360–69.

R. Hinchcliff, Testimony before the Committee on Surface Transportation, Committee on Commerce, U.S. Senate, May 12, 1972, *The Case Against Deregulation* (American Trucking Assn.).

D. G. Kendall, "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains," *Annals Math. Statist.*, Sept. 1953, *24*, 338–54.

R. A. Meyer, "Monopoly Pricing and Capacity Choice Under Uncertainty," *Amer. Econ Rev.*, June 1975, *65*, 326–37.

H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, *60*, 693–705.

J. E. Olson, "Price Discrimination by Regulated Motor Carriers," *Amer. Econ. Rev.*, June 1972, *62*, 395–402.

S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation In Pure Competition," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 34–55.

G. J. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1962, *69*, 213–25.

P. L. Swan, "Quality of Consumption Goods," *Amer. Econ. Rev.*, Sept. 1970, *60*, 884–94.

L. G. Telser, "Searching for the Lowest Price," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 40–49.

An Empirical Inquiry on the Short-Run Dynamics of Output and Prices

By ROQUE B. FERNANDEZ*

The purpose of this paper is to study the short-run relationship between output and inflation in the context of a macroeconomic model. Although a considerable number of economists have studied this subject, mainly from the point of view of the Phillips curve theory, most of them have made use of *ad hoc* hypotheses regarding the process through which expectations are formed. Other economists (for example, Robert Lucas, and Thomas Sargent and Neil Wallace) have studied the same subject and have postulated a rational expectation hypothesis for analyzing the short-run tradeoff between inflation and output and for testing the "natural rate" hypothesis.

The analysis of this paper is similar to that of Lucas, and Sargent and Wallace. In fact, the analysis in Section I starts with the assumption that the model previously postulated by Sargent and Wallace is an appropriate theoretical framework for analyzing the short-run relationship between prices and output. That section presents the underlying model as well as some of its main limitations and implications.

In Section II a summary of the results of the structural analysis of the model is presented as well as the estimation procedure followed in order to obtain the estimates for the structural equations of the system. Section III presents some estimates based on the available information for Argentina and Brazil. Finally, concluding remarks are made in Section IV.

I. The Macroeconometric Model

The model to be analyzed in this section is a standard macroeconomic model in which expectations will be assumed to be "rational" in the sense of John Muth. This assumption was incorporated in similar models by Sargent, and Sargent and Wallace. In this paper some modifications are introduced in order to arrive at a directly estimable relationship for a short-run output-inflation tradeoff.

The model consists of the following three equations:

Aggregate supply:

$$(1)\quad y_t = y_{n,t} + a(p_t - {}_tp^*_{t-1}) + k(y_{t-1} - y_{n,t-1}) + u_{1t} \qquad a > 0$$

Aggregate demand:

$$(2)\quad y_t = y_{n,t} + g + c(r_t - ({}_{t+1}p^*_{t-1} - {}_tp^*_{t-1})) + u_{2t} \qquad c < 0, \; g > 0$$

Portfolio balance:

$$(3)\quad \phi m_t = p_t + y_t + br_t + u_{3t} \qquad -\infty < b < 0$$

In these equations y_t, p_t, and m_t are the natural logarithms of real income, the price level, and the nominal stock of money; g is a constant; and the $u_{i,t}$ $(i = 1,2,3)$ are disturbance terms with zero means that may be serially and contemporaneously correlated. The variable $y_{n,t}$ is a measure of normal productive capacity that will be represented by the trend in real output in the empirical application of the model. There-

*International Monetary Fund. The present paper was completed while I was at the Latin American Institute for Economic and Social Planning (ILPES, United Nations, Chile). I am grateful to Arnold Zellner, Robert Barro, Arnold Harberger, Milton Friedman, Robert Lucas, and other members of the Money and Banking Workshop of the University of Chicago for comments on an earlier draft. I also want to thank James Hanson and Luis E. Rosas of ILPES for valuable discussions. This paper draws on material from my doctoral dissertation. A preliminary version was presented at the Conference on Planning and Short-Term Macroeconomic Policy in Latin America, October-November 1975, Panama, sponsored by ILPES, the Ministerio de Planificacion y Politica Economica de Panama and the National Bureau of Economic Research.

fore, $y_t - y_{n,t} = y_{c,t}$ represents cyclical or "detrended" output. The variable ${}_{t+1}p_t^*$ represents the public's expectation at time t of logarithm of the price level expected to prevail at t + 1. The variable r_t is the nominal rate of interest.

Equation (1) is an aggregate supply equation relating detrended output to the gap between current price level and the public's prior expectation of the current price level. In this equation lagged detrended output indicates that deviations of aggregate supply from normal capacity may display some persistence. This same equation was postulated and used by Lucas.

Equation (2) is an aggregate demand equation which relates the deviation of aggregate demand to the real rate of interest, which in turn is represented by the nominal rate of interest minus the expected rate of inflation. This equation, used by Sargent and Wallace, differs from the one used by Sargent in its definition of the real rate of interest. Sargent uses the usual definition, that is, $r_t - ({}_{t+1}p_{t-1}^* - p_t)$. However the definition stated in equation (2) seems more plausible in that demanders at time t do not observe p_t, so they have to anticipate it. Notice that this is not inconsistent with the inclusion of p_t in (1), because equation (1) does not imply that suppliers observe the general price level, but is derived from an aggregation of the response of individual suppliers, who only observe the prices in their markets and not the general price level (see Lucas, pp. 327-28).

Some limitations of equation (2) are stated in Sargent as follows:

> [A]n important thing about equation (2) is that it excludes as arguments both the money supply and the price level, ... This amounts to ruling out direct real balance effects on aggregate demand. It also amounts to ignoring the expected rate of real capital gains on cash holdings as a component of the disposable income terms that belong in the expenditure schedules that underlie equation (2). Ignoring these things is usual in macroeconometric work. [p. 435, fn. 15]

Another aspect of this model is the lack of symmetry between equations (1) and (2). That is, only suppliers have explicit misperceptions of prices and only demanders have explicit responses to change in the real rate of interest. However, the effect of the neglected variables in each equation could be captured implicitly if they induced some stable stochastic process in the error terms.

Equation (3) is a demand for money relationship with unit real income elasticity—an assumption that is not crucial and will be relaxed. It summarizes the condition for portfolio equilibrium. In other words, when equation (3) is satisfied, owners of bonds and equities are satisfied with the division of their portfolio between money (assumed to be exogenous) on the one hand, and bonds and equities on the other hand. The ϕ is a polynomial in the lag operator (that is, $\phi = \phi_0 + \phi_1 L + \phi_2 L^2 + \ldots,$ where $\phi_0 + \phi_1 + \ldots = 1$) introduced in an effort to capture the effects of lagged changes of m_t on nominal income.[1] The degree of this polynomial will be determined empirically.

On purely theoretical grounds, there is not a strong justification for the existence of a lagged response of nominal income to changes in the quantity of money. However, the existence of lags is confirmed in most of the empirical work that relates money and prices.

The workings of the model can easily be illustrated. Consider for the moment only equations (2) and (3) that resemble the sim-

[1]Equation (3) can be derived from the simple quantity theory, that is

$$\text{(a)} \qquad Y_t \cdot \frac{1}{V_t} = M_t$$

Now in order to capture the lagged effect of M_t on the left-hand side, we have to specify something like

$$Y_t \cdot \frac{1}{V_t} = f(M_t, M_{t-1}, M_{t-2}, \ldots)$$

and a specific construction is

$$\text{(b)} \qquad Y_t \cdot \frac{1}{V_t} = \exp(\phi \, ln \, M_t)$$

where ϕ is polynomial in the lag operator (notice that if $\phi = 1$ we get (a)). Assuming $(1/V_t) = \exp(br_t)$ and taking *logs* on both sides of (b) we get equation (3) of the text.

ple textbook *IS-LM* model and leave aside the problem of how expectations are formed. Equation (2) corresponds to equilibrium in the real sector and relates real income to the real rate of interest, that is, the *IS* curve. Equation (3) refers to equilibrium in the monetary sector usually represented by the *LM* curve. The system formed by equation (2) and (3) is not determined because we have two equations and three endogenous variables, y, p, and r. This problem is solved in the standard textbook analysis by assuming either that prices are rigid so the shifting of the *IS* and *LM* curves affects real output (this would be the simplest Keynesian model), or assuming that the economy is at full employment so the shiftings of the *IS* and *LM* curves would only affect prices (this would be the simplest quantity theory approach). In our case we solve the problem of joint determination of prices and output through equation (1) which in turn adds an important feature to the model: namely the relationship between output and price misperceptions.

To complete the above model we should specify how expectations are formed. This is a delicate matter. It is customary to postulate different *ad hoc* hypotheses about the formation of expectations. The most popular version is the Cagan hypothesis of adaptive expectations, although the explanation for its use was confined to the fact that adaptive expectations seemed reasonable and proved useful in explaining data. The hypothesis of rational expectations used in this paper follows John Muth's proposal that expectations are informed predictions of future events based on the available information and the relevant economic theory. This has one strong implication—the economist who is modelling an economy does not have a superior knowledge of the "reality." This view is in turn confirmed by the fact that samples of predictions "are more accurate than naive models and as accurate as elaborate equation systems" (Muth, p. 316). Thus, our model is completed with the following equations:

$$(4) \qquad {}_tp^*_{t-1} = Ep_t$$

$$(5) \qquad {}_{t+1}p^*_{t-1} = Ep_{t+1}$$

where Ep_t is the conditional mathematical expectation of p_t formed using the model and all the information assumed to be available as of the end of period t − 1 (hereafter the E operator will always be conditional on the information available as of the end of period t − 1).

After some algebraic manipulations of the model, the following two equations for expected prices can be obtained:

$$(6) \quad Ep_t = [1/(1-b)]\sum_{j=0}^{\infty} \cdot [1/(1-b^{-1})]^j[E\phi m_{t+j} - y_{n,t+j}] + [J_3/(1-J_0)]\sum_{j=0}^{\infty} \cdot [k/(1-b^{-1})]^j y_{c,t-1} + c_0$$

$$(7) \quad Ep_{t+1} = [1/(1-b)]\sum_{j=0}^{\infty} \cdot [1/(1-b^{-1})]^{j+1}[E\phi m_{t+j+1} - y_{n,t+j+1}] + [J_3/(1-J_0)]\sum_{j=0}^{\infty} \cdot [k/(1-b^{-1})]^{j+1} y_{c,t-1} + c_0$$

In these equations J_0, J_3, and c_0 are constants that are complicated functions of the structural parameters of the model.[2]

From these equations, it is easy to illustrate the process of formation of expectations. Assume for a moment that $b = 0$ (that is, that the interest elasticity of the demand for money is zero). Then, after taking first differences (D operator), equation (6) can be reduced to

$$(8) \quad EDp_t = \phi_0 EDm_t - (\beta + kDy_{c,t-1}) + \phi_1 Dm_{t-1} + \phi_2 Dm_{t-2} + \ldots$$

[2]The algebra for obtaining equations (6) and (7) is similar to the methodology developed by Sargent and Wallace. For a complete derivation see the author, Appendix A.

$$J_0 = [a + c/(1 + cb^{-1})]/\theta$$
$$J_3 = -k/\theta$$
$$\theta = a + cb^{-1}/(1 + cb^{-1})$$
$$c_0 = [g/c(b^{-1} - 1)]\sum_{j=0}^{\infty}[1/(1 - b^{-1})]^j$$

where β is the slope coefficient of the trend in real output and $J_3/(1 - J_0) = -k$ when $b = 0$. Equation (8) clearly shows that the expected rate of change of prices depends upon the expected rate of change in the money supply in period t; the natural rate of growth in output (β), a term in the cyclical component of output in t − 1; and past rates of change of the money supply. If $b \neq 0$ the results are not far from the quantity theory in expectation form, although the algebraic expression representing the expectation formation process is more complicated.

The money supply process which forms the basis on which the public makes its forecasts of the future path of m_{t+j} is of particular relevance.[3] The empirical analysis of Section III considers two processes as determining the money supply: The first is an *ARIMA* process that in its "inverted form" is

$$(9)\quad Dm_t = \pi_1 Dm_{t-1} + \pi_2 Dm_{t-2} + \pi_3 Dm_{t-3} + \ldots v_{3t}$$

where the π's are parameters. The second process will be a model of the form

$$(10)\quad Dm_t = \pi_1' Dm_{t-1} + \pi_2' z_t + u_t$$

where π_1' can be a parameter or a polynomial in the lag operator and π_2' can be a row vector of parameters or a row vector of polynomials in the lag operator while z_t is a column vector of predetermined variables.

The empirical tests will not be carried out directly in the form of equations (9) and (10) but indirectly through the transfer functions of the next section.

II. Towards an Empirical Test of the Model

In this section I outline the method used to estimate the structural equations of the

[3]In searching for a process determining the money supply we can choose either to postulate a model for the money supply by relating it to a set of "predetermined variables" relative to the model (1)-(3) (so m_t still remains as if it were exogenous or determined outside of the system (1)-(3)) or, we can identify a Box-Jenkins *ARIMA* process. It has been customary in the economics profession to call these models "naive models" because of the rather simple structure by which only past values of a variable are used to predict future values of the same variable. However, it has recently been shown (see Zellner and Palm) that these models might not be naive at all. Indeed these models (the *ARIMA* models) represent the "final form" for a variable implied by a highly sophisticated model. I will briefly illustrate this point with a model for the nominal money supply. Let us assume that in a given country the money supply is generated by the following relationship

$$(a)\quad Dm_t = c_1 + a_1 Dm_{t-1} + b_1 Dg_t + e_1 Dx_t + v_t$$

where g_t could be the federal budget relative to lagged *GNP*, x_t could be the lagged balance-of-payments surplus relative to *GNP*, and v_t an error term stochastically independent of the errors in the structural equations. In our case a_1, b_1, and e_1 are assumed to be constants for simplicity, but in a more general analysis, we could assume a_1, b_1, and e_1 to be polynomials in the lag operator. Now we shall show that equation (a) implies a final equation for m_t that is in the form of an *ARIMA* process. Equation (a) can be written as

$$(b)\quad (1 - a_1 L) Dm_t = c_1 + b_1 Dg_t + e_1 Dx_t + v_t$$

Let the predetermined variables g_t and x_t follow any process over time. That is, both could follow a random walk or one could follow a random walk and the other a given *ARIMA* process, etc. To illustrate the problem at hand we will assume that

$$Dg_t: ARIMA\,(2, 1) \quad \text{or} \quad \phi(2)\, Dg_t = \phi(1) v_{1t}$$

$$Dx_t: \text{random walk} \quad \text{or} \quad \phi(0)\, Dx_t = \phi(0) v_{2t}$$

where the v's are stochastically independent of the disturbances in the structural equations. Multiplying both sides of (b) by $\phi(2)\,\phi(0)$ we have

$$(c)\quad (1 - a_1 L)\phi(2)\phi(0) Dm_t = \phi(2)\phi(0)c + b\phi(0)\theta(1)v_{1t} + e\phi(2)\theta(0)v_{2t} + \phi(2)\phi(0)v_t$$

In this last expression we notice that we have obtained an *ARIMA* (3, 2) process (if no cancellation occurs) for m_t as implied by equation (a) and the assumption for the predetermined variables g_t and x_t. This clearly illustrates that if we obtain the process *ARIMA* (3, 2) for m_t this is not a naive model at all, but on the contrary it could be reflecting the "true" model governing the behavior of the money supply. The above discussion demonstrates that we cannot talk about "alternative models" when we evaluate a model of the sort of equation (a) with respect to a model like (c) because this could be the final form of (a). While we have considered it appropriate to check empirically the *ARIMA* hypothesis for m_t as well as a model of the sort implied by equation (a), no further attention is given to the "theory of the money supply" that underlies our hypothesis of the money supply process, a subject that goes beyond the scope of this paper.

model. Two points are jointly developed, one is the computation of expected prices and the other is the endogeneity of p_t that precludes the straightforward estimation of equation (1) using ordinary least squares.[4]

At the estimation stage we shall concentrate on equations (1) and (3). The main problem with equation (2) is the variable r_t, for which we do not have data for some countries (for example, Argentina and Brazil). This problem is eliminated in equation (3) because it is assumed that variation in r_t is dominated by variation of the expected rate of inflation, and the public's forecast of the rate of inflation (based on information available at t − 1) is used as a proxy for r_t. It is obvious that this substitution cannot be made in equation (2): the term $(r_t - ({}_{t+1}p^*_{t-1} - {}_tp^*_{t-1}))$ would vanish when r_t is replaced by ${}_{t+1}p^*_{t-1} - {}_tp^*_{t-1}$. Nevertheless, the system formed by equations (1) and (3) is perfectly determined when a proxy is used for r_t, let us say, $r'_t = Dp^*_t = {}_{t+1}p^*_{t-1} - {}_tp^*_{t-1}$.

Let us first consider equation (1). We know that a direct estimation of this equation is not possible because p_t and y_t are jointly determined and ${}_tp^*_{t-1}$ is not observable. Thus, in this section our objective is to obtain an estimable relationship in place of (1), making use of the relationships previously developed.

In equation (6) we obtained an expression for the expectations formation process in which the expected *log* of the price level in period t was determined by the *log* of the money supply expected to prevail in period t, by the trend in the *log* of real output, and by the detrended *log* in real output in period t − 1. Clearly, the actual *log* of the price level differs from the expected value by a random component, say u_{4t}, so we can write

$$(11) \qquad p_t = Ep_t + u_{4t}$$

Then our hypothesis implies that the expected p_t is computed as if the public attempted to obtain an optimal unbiased forecast of p_t using equation (6). Combining (11) and (6) we can write

$$(11') \quad p_t = [1/(1-b)]\sum_{j=0}^{\infty}[1/(1-b^{-1})]^j \cdot [E\phi m_{t+j} - y_{n,t+j}] + [J_3/(1-J_0)] \cdot \sum_{j=0}^{\infty}[k/(1-b^{-1})]^j y_{c,t-1} + c_0 + u_{4t}$$

In (11′) we have a term in $E\phi m_{t+j}$. Developing this term for $j = 0, 1, \ldots$, taking expectations and recalling that

$\phi = \phi_0 + \phi_1 L + \phi_2 L^2 + \ldots$, we have

$$E\phi m_{t+j} = E\phi_0 m_t + \phi_1 m_{t-1} + \phi_2 m_{t-2} + \ldots \quad j = 0$$

$$E\phi m_{t+j} = E\phi_0 m_{t+1} + E\phi_1 m_t + \phi_2 m_{t-1} + \ldots \quad j = 1$$

$$E\phi m_{t+j} = E\phi_0 m_{t+2} + E\phi_1 m_{t+1} + E\phi_2 m_t + \phi_3 m_{t-1} \ldots \quad j = 2$$

Recall that the E operator is conditional on the information in period t − 1, so $Em_{t-1} = m_{t-1}$ and so on for periods before period t − 1. Now provided that we use the process (9) to obtain Em_{t+j}, $j = 1, 2, \ldots$, we notice that the forecasts of m_t are obtained through linear combinations of m_{t-1}, m_{t-2}, m_{t-3}, ... These linear combinations should be combined with the other terms in $m_{t-1}, m_{t-2}, m_{t-3}, \ldots$ that appear because of

[4]Some testable implications of the model can be derived from a structural analysis of the system. This analysis, following the method suggested in Zellner and Palm, is presented in the author where the final equations of the system (1)–(5) were derived and compared with the data. Also in that work a variant of the system is analyzed in which an adaptative expectation hypothesis was used for prices. This version was incompatible with the available information for both Argentina and Brazil while the rational expectations version of the model (system (1)–(5)) was compatible under certain conditions. As shown in the author, given a system of structural equations, we can work out the "final equations" for the variables of the system. These are in the form of *ARIMA* processes. On the other hand, we can identify the *ARIMA* processes for the variables using the available information on each variable. If the structural equations of the model are correct, the final equations derived for each endogenous variable should have the same structure that the *ARIMA* processes identified for those variables from the available information. If this is the case, we say that the model is *compatible* with the available information.

the lagged response of prices to changes in m_t, and with $y_{n,t+j}, j = 1, 2, \ldots$ and $y_{c,t-1}$ to forecast p_t. Then we can rearrange the terms in $m_{t-1}, m_{t-2}, m_{t-3}, \ldots$, and rewrite (11′) as

$$(11'') \quad p_t = v(L)Lm_t - (1/(1-b)) \cdot \sum_{j=0}^{\infty}(1/(1-b^{-1}))^j y_{n,t+j} + (J_3/(1-J_0)) \sum_{j=0}^{\infty} (k/(1-b^{-1}))^j y_{c,t-1} + c_0 + u_{4t}$$

where $v(L)$ is a polynomial in the lag operator ($Lx_t = x_{t-1}$) capturing the effect of all the linear combinations on past values of m_t on p_t.

The first difference form of this equation is

$$(11''') \quad Dp_t = v(L)LDm_t + h_0 Dy_{c,t-1} + c + u_{5t}$$

where c accounts for the term in $y_{n,t+j}$ after differencing (recall that $y_{n,t}$ is a trend and differencing it yields the slope coefficient of the trend line) and h_0 represents the coefficient of $y_{c,t-1}$.

To estimate (11‴) we have to consider the problem of collinearity, especially in the case of quarterly data, where a reasonable lag of two years would imply that m_t should be lagged eight times. A way of dealing with equation (11‴) is to consider it to be a multiple input transfer function.[5] The transfer function form of equation (11‴) can be parsimoniously (in terms of the number of parameters) represented by

$$(12) \quad Dp_t = \frac{w_1(L)}{\alpha_1(L)} LDm_t + \frac{w_2(L)}{\alpha_2(L)} LDy_{c,t} + \frac{\theta(L)}{\phi(L)} u_t + c$$

The estimation of (12) can be done using the Marquardt algorithm, and the forecasts made using the estimated version of (12) are minimum mean square error forecasts.[6] Then, from (12) we can obtain a series of "expected prices." We now need to compute a series of "actual prices." Recall that in equation (1) we cannot compute the difference $p_t - {}_tp^*_{t-1}$ and estimate that relationship, because in our model both p_t and $y_{c,t}$ are endogenous. Also by straightforward algebra (using (6), (7) and (9)), we obtain

$$(13) \quad Dp_t = \frac{w'_1(L)}{\alpha'_1(L)} Dm_t + \frac{w'_2(L)}{\alpha'_2(L)} LDy_{c,t} + \frac{\theta'(L)}{\phi'(L)} u'_t + c'$$

where the meaning of the notation is the same as in equation (12). Notice that the main difference between (12) and (13) is that in (12) m_t appears lagged one period.

Now recall that equation (9) represents the hypothesis that the money supply follows an *ARIMA* process. If we use assumption (10) for the money supply, then equations (12) and (13) should be extended to include terms in the components of z_t.

Although the algebraic analysis is rather long, its intuitive interpretation is quite straightforward. The rational expectation feature of the model implies that the public forms their expectations using the information available at the end of period t − 1. In forming these expectations, the money supply expected to prevail in future periods is important, and it is assumed that the public forecasts future values of m_t by considering the history of m_t available at t − 1 (as well as other variables if (10) is used). But the history of m_t is not only relevant for forecasting future values; the recent past values of m_t also directly affect the price level because of the lagged response of prices to changes in the money supply. This is also considered in the expectations formation process. Equation (12) is oriented to capture this process.

[5]The analysis of transfer functions can be found in chapters 10 and 11 of Box and Jenkins. A derivation of transfer functions, for a simultaneous equation model different from the one presented in this paper, can be found in Zellner and Palm.

[6]The computer programs for estimation of transfer functions were provided to the author by Charles Nelson.

Equation (13), although very similar to (12), is quite different. It is a reduced form for p_t, implied by the system (1)-(5) and the assumption in (9) or (10) for the money supply. In (13), m_t directly affects the price level. The economy as a whole need not forecast m_t; it is an exogenous variable determined by monetary authorities in period t which will have an immediate effect on p_t.

The fitted values for p_t from (13) will be introduced in (1) in place of p_t and the fitted values of (12) will be introduced in (1) in place of ${}_tp^*_{t-1}$ in order to estimate equation (1).

Now consider equation (3). This equation assumes that the real income elasticity of demand for money is one. This assumption need not be maintained since all the algebraic expressions that we obtained before can be rearranged to include an additional parameter (the real income elasticity of the demand for money). Hereafter we will relax this assumption by writing (3) as $Y_t = \phi m_t - br'_t - u_{3t}$, where $Y_t \equiv p_t + iy_t$; i representing the real income elasticity of demand for money. It should be noticed that if $i = 1$ then Y_t is the *log* of nominal income. Then the system (1)-(3) can be interpreted as follows: equation (3) determines nominal income and equation (1) determines the division of nominal income between changes in prices and changes in output.

For analyzing the cases in which $i \neq 1$ we will evaluate the results for three cases: $i = 0.5$, 1.5, and 2. An attempt was made to estimate i using an *ARIMA* process as the the instrumental variable for Dy_t. The results, however, were not reliable because y_t behaves almost like a random walk. At the estimation stage, equation (3) will be expressed in the form of a transfer function with all variables in first difference. The transfer function form will allow us to estimate the lag structure induced by the polynomial lag operator ϕ.

III. Empirical Results

In this section we proceed to test and estimate the model presented in Section II and III with the available data for Argentina and Brazil. First, we construct a series of expected prices on the basis of the results obtained in fitting equation (12). Second, we construct a series of actual prices from the reduced form for prices—equation (13) (recall that this step is necessary in order to avoid the problem of simultaneity in estimating equation (1)). "Actual prices" minus expected prices give us the misperceptions of prices required for estimating equation (1). Finally, we estimate equation (3) under different assumptions with respect to the real income elasticity of the demand for money, using a proxy for the interest rate.

A. *The Data*

All the data for Argentina were obtained from *International Financial Statistics (IFS)* and *Boletin de Estadistica*. They include quarterly data for the index of industrial production, wholesale prices, currency and demand deposits, wages set in collective bargaining and the balance of trade (all seasonally adjusted by the method of moving averages). The observations relate to the period 1956-I to 1973-II (this period was chosen in order to base the analysis on the maximum number of observations available for the index of industrial production.[7]

The *log* of the index of industrial production for Argentina was detrended splitting the data into two parts: from 1956-I to 1962-IV and from 1963-I to 1973-II. This was done because there is no trend in real output in the first period, and if a single trend line were fitted to the whole period we would lose most of the cyclical fluctuations.[8]

[7]The index of industrial production is used for Argentina as a proxy for real income because it is more reliable and comprehensive than existing series of real output. For Brazil the only available information corresponds to real output.

[8]As a matter of fact, this was exactly the procedure originally followed. The procedure was abandoned because the detrended output obtained in this manner showed an initial period in which output was mostly above the trend, a second period of almost "seven years" in which output was below the trend, and a third

TABLE 1—ESTIMATED TRANSFER FUNCTION FOR EXPECTED PRICES
(Estimates of Equation (12))

Model	Estimates of the *AR* and *MA* parts	Residual Sum of Squares (*RSS*); Degrees of Freedom (*DF*); *RSS/DF*; Adj. R^2	Estimate of $Dy_{c,t-1}$	Dummy or Constant	Wages	Balance of Trade
	Argentina 1956-I-1973-II					
(1)	Dm_{t-1}: $0.492/(1 - 1.041L + 0.785L^2)$	0.122293	−0.006	0.011		
	(0.129) (0.082) (0.087)	59	(0.094)	(0.012)		
	u_t: $1/(1 - 0.402L - 0.252L^2)$	0.00207				
	(0.133) (0.130)	.47				
(2)	Dm_{t-1}: $(0.693 - 0.317L)/(1 - 1.112L - 0.719L^2)$	0.117639	−0.060	0.011		
	(0.166) (0.221) (0.135) (0.102)	57	(0.104)	(0.013)		
	u_t: $1/(1 - 0.376L - 0.271L^2)$	0.00206				
	(0.135) (0.131)	.48				
(3)	Dm_{t-1}: $0.613/(1 - 1.009L + 0.741L^2)$	0.116154	−0.022	0.006	0.044	−0.004
	(0.146) (0.099) (0.098)	59	(0.091)	(0.009)	(0.081)	(0.001)
	u_t: $1/(1 - 0.436L)$	0.00197				
	(0.124)	.49				
	Brazil 1955-I-1971-IV					
(1)	Dm_{t-1}: $0.188/(1 - 0.723L)$	0.066609	−0.009	0.016		
	(0.122) (0.174)	60	(0.077)	(0.024)		
	u_t: $1/(1 - 0.520L)$	0.00111				
	(0.116)	.46				
(2)	Dm_{t-1}: $0.199/(1 - 1.388L + 0.556L^2)$	0.065183	−0.012	0.013		
	(0.099) (0.662) (0.578)	59	(0.078)	(0.023)		
	u_t: $1/(1 - 0.508L)$	0.00110				
	(0.116)	.46				

Notes: The IMF was the main source of monetary, price, and industrial production data. For Brazil, real output was obtained from Goncalves. The terms *AR* and *MA* represent the autoregressive and moving average parts, respectively, of the rational polynomials. Large sample standard errors of the parameters are beneath them in parentheses.

The data for Brazil were obtained from two sources: *International Financial Statistics (IFS)* and Goncalves. From the *IFS* I obtained the series of wholesale prices (excluding coffee) and currency and demand deposits. From Goncalves I obtained a series of quarterly real output. All the observations relate the period 1955-I to 1971-IV (this period was chosen in order to base the analysis on the maximum number of observations available for real output). The data were seasonally adjusted by the method of moving averages.

All the variables were expressed in first difference of *logs* prior to estimation except the balance of trade. This variable was computed as the *log* of exports minus the *log* of imports because of the impossibility of taking *log* of a negative number in the case of trade deficits.

period where output was above the trend. A detailed explanation about some institutional aspects that could explain the difference in the trend of real output above mentioned can be found in the author, pp. 36–39.

B. *Estimates of the Transfer Function for Expected Prices*

In Table 1 the estimates obtained for equation (12), are presented. This is the expression that determines expected prices.[9] These models have been selected from a larger number of models with different lag structures and different error terms. The selection has been carried out using the likelihood ratio test proposed by Arnold Zellner and Franz Palm.

Models (1) and (2) for Argentina assume that the money supply follows a process as represented by equation (9), and model (3) considers the assumption implied by (10). In model (3) we have computed the transfer function with wages and balance of trade as input variables. We see from Table 1 that in the case of Argentina there is a slight reduction in the residual variance (RSS/DF) and a small increase in the adjusted R^2 when passing from model (1) or (2) to model (3).[10] At the bottom of the table we present the results obtained for Brazil where an insignificant reduction in the residual variance is observed when we go from the simple lag structure of model (1) to the more complex lag structure of model (2).

C. *Estimates of the Reduced Form for Prices*

Table 2 shows the estimates of the transfer functions for prices (that is, equation (13)). Here again, for Argentina models, (1) and (2) incorporate assumption (9) for the money supply while model (3) incorporates assumption (10). In both Tables 1 and 2 the coefficient of the Balance of Trade variable is significantly different from zero at the 5 percent level. Only in Table 2 does the Balance of Trade variable have an estimated coefficient with the positive sign that the theory predicts.

In the case of Brazil we observe again that no appreciable reduction in the residual variance is obtained in going from the simple lag structure of model (1) to the more complex lag structure of model (2).

In both Tables 1 and 2 the estimates for the variables $Dy_{c,t-1}$ and dummy or constant are small numbers, not significantly different from zero. This is not in contrast with the theoretical model because these parameters can indeed be close to zero (see equation (11″)).

D. *Estimates of the Aggregate Supply*

Recall that Table 1 provides the estimates of equation (12) which in turn allow us to obtain a series of "expected prices" needed to estimate equation (1) of our original model. By the same token, equation (13) whose estimates are given in Table 2 provides us with a series of "actual prices" to estimate equation (1). Then the next step is to compute a one-step ahead forecast from (12) that would give us a proxy variable for ${}_tp^*_{t-1}$; similarly a one-step ahead forecast from (13) would give us a proxy for p_t. The difference between the proxy of p_t and the proxy for ${}_tp^*_{t-1}$ is introduced in equation (1) in place of $(p_t - {}_tp^*_{t-1})$, and the estimation of this equation provides us with an estimate of the slope coefficient of our short-run Phillips equation. Table 3 shows the results obtained by this procedure and indicates the different models used for forecasting prices and reduced forms used for prices. On testing the significance of the Phillips parameter for Argentina using a two-tailed test we notice that at the 5 percent level only regression (5) shows an estimate significantly different from zero. Using a one-tailed test (the alternative hypothesis is that the parameter is greater than zero),

[9]The column headed "dummy" corresponds to the constant c in equation (12). The dummy appears in the empirical results for Argentina because the constant c is a term in the slope coefficient of the trend line for output. When we split the data into two periods, with a different slope coefficient in each period, a dummy with value of one from 1956-I to 1962-IV and two from 1963-I to 1973-II was incorporated in the transfer functions takes into account the correction for degrees in trend.

[10]The adjusted R^2 reported in the tables for transfer functions takes into account the correction for degrees of freedom. That is,

$$1 - R^2\text{adj} = \frac{n-1}{n-k}(1 - R^2)$$

TABLE 2—ESTIMATED TRANSFER FUNCTIONS FOR PRICES
(Estimates of Equation (13))

Model	Estimates of the AR and MA parts	Residual Sum of Squares (RSS); Degrees of Freedom (DF); RSS/DF; Adj. R^2	Estimate of $Dy_{c,t-1}$	Dummy or Constant	Wages	Balance of Trade
	Argentina 1956-I–1973-II					
(1)	Dm_t: $0.422/(1 - 1.216L + 0.773L^2)$	0.125758	0.049	0.004		
	(0.112) (0.102) (0.103)	59	(0.095)	(0.012)		
	u_t: $1/(1 - 0.400L - 0.216L^2)$	0.00213				
	(0.133) (0.132)	.45				
(2)	Dm_t: $(0.224 + 0.346L)/(1 - 1.073L + 0.763L^2)$	0.116639	0.030	0.004		
	(0.160) (0.202) (0.104) (0.091)	57	(0.103)	(0.012)		
	u_t: $1/(1 - 0.365L - 0.280L^2)$	0.00205				
	(0.135) (0.132)	.48				
(3)	Dm_t: $(0.207 + 0.435L)/(1 - 1.054L + 0.749L^2)$	0.113562	0.010	−0.003	0.018	0.003
	(0.180) (0.229) (0.116) (0.098)	57	(0.095)	(0.010)	(0.084)	(0.001)
	u_t: $1/(1 - 0.438L)$	0.00199				
	(0.124)	.49				
	Brazil 1955-I–1971-IV					
(1)	Dm_t: $0.212/(1 - 0.716L)$	0.064908	−0.004	0.010		
	(0.117) (0.155)	61	(0.049)	(0.023)		
	u_t: $1/(1 - 0.511L)$	0.00106				
	(0.113)	.47				
(2)	Dm_t: $0.119/(1 - 1.389 + 0.546L^2)$	0.064130	−0.005	0.008		
	(0.190) (0.630) (0.611)	60	(0.076)	(0.022)		
	u_t: $1/(1 - 0.498L)$	0.00107				
	(0.115)	.47				

Notes: See Table 1.

estimates of regressions (3), (4), and (5) provide evidence for rejecting the null hypothesis at the 5 percent level of significance. In all the cases the Box and Pierce Q-statistic is in favor of rejecting the hypothesis of autocorrelation of residuals.[11]

It is convenient at this stage to take a closer look at the estimates of Table 3. Recall that the estimate of parameter a is an estimate of the slope of the Phillips curve. Our results for Argentina indicate that there is some evidence in favor of a short-run tradeoff between inflation and output given by the 95 percent confidence intervals for the estimates of regressions (3), (4), and (5). These are (−0.089, 0.991), (0.0, 1,384), and (0.001, 2,268), respectively. However the short-run tradeoff that we have found

[11] The Q-statistic is calculated from the first K autocorrelations r_k $(k = 1, 2, \ldots, K)$. If the fitted model is appropriate,

$$Q(K) = n \sum_{k=1}^{K} r_k^2$$

is approximately distributed as $\chi^2(k - p - q)$. If the model is wrong the value of Q will be inflated. In the case of Table 3, $p = q = 0$ because there are no autoregressive or moving average parameters in the noise model.

TABLE 3—ESTIMATES OF THE AGGREGATE SUPPLY EQUATION
(Estimates of Equation (1))

Model for Reduced Form	Model for Expected Prices	a	k	Adjusted R^2	Q-Statistics
		Argentina 1956-I-1973-II			
(1) M_2	M_2	0.877 (0.594)	0.564 (0.102)	.35	15.3
(2) M_2	M_1	0.647 (0.578)	0.574 (0.102)	.34	16.5
(3) M_1	M_1	0.401 (0.295)	0.575 (0.102)	.35	15.7
(4) M_1	M_2	0.692 (0.346)	0.569 (0.101)	.36	14.2
(5) M_3	M_3	1.140 (0.564)	0.778 (0.102)	.35	15.1
		Brazil 1955-I-1971-IV			
(1) M_1	M_1	−0.272 (0.919)	0.664 (0.095)	.430	15.1
(2) M_1	M_2	−0.492 (0.750)	0.660 (0.095)	.433	16.0
(3) M_2	M_1	0.842 (1.328)	0.657 (0.095)	.433	14.5
(4) M_2	M_2	−0.140 (1.390)	0.665 (0.095)	.429	14.9
(5) Actual Prices	M_1	0.168 (0.178)	0.634 (0.099)	.437	15.8
(6) Actual Prices	M_2	0.153 (0.172)	0.638 (0.099)	.436	15.5

Note: Chi-Square values from table. $\chi^2(24) = 33.2$ 0.10 level of significance
$\chi^2(24) = 36.4$ 0.05 level of significance

The IMF was the main source of industrial production data for Argentina and Goncalves for Brazil. Standard errors of the parameters are in parentheses. The models used to represent prices and expected prices are symbolized in this table with the letter M and a subindex. Thus, M_2 in the first column means that model (2) of Table 2 is being used to represent actual prices in the aggregate supply equation.

does not contradict the natural rate hypothesis of Milton Friedman. As equation (1) indicates, if prices are anticipated correctly, output will remain in its long-run trend (or "natural" level).

In the case of Brazil we notice that in equations (1), (2), and (4) the estimate of a is negative although not significantly different from zero at the 0.05 level in a two-tailed test. Model (3) presents the right sign, but its a estimate has a large standard error that makes it not significantly different from zero. In all the cases the value of the Q-statistics favor rejection of the hypothesis of autocorrelation in the residuals.

In order to compare our results with other results obtained for Brazil by Goncalves, I estimated the last two models of Table 3, where the actual prices were included instead of the forecast of the reduced form for prices. Goncalves did a similar estimation under the assumption that the price level was exogenously determined (mainly due to strongly enforced price controls in most of his period of analysis). He worked with the period 1959–69 and used another hypothesis for expectations formation. His results provide an estimate of a equal to 0.41 (standard errors are not reported in his work). When a dummy variable for the period 1961-I to 1963-II is included to capture the effect of price controls, his results shows a equal to 0.27. It

should be noted that this last result, obtained by Goncalves, is quite close to the estimates in models (5) and (6) of Table 3. From our results for Brazil we must conclude that the empirical evidence argues against a stable tradeoff between output and inflation even in the short run.[12]

E. *Estimates of the Transfer Function for Nominal Income*

Now we proceed to the estimation of equation (3) of our original model. Recall that in this equation we are using nominal income as the dependent variable when the real income elasticity of the demand for money is assumed equal to one; and we are using as dependent variable the term $p + iy$ where i is the real income elasticity for all the cases in which it is assumed that $i \neq 1$. Also, we are using a one-step ahead forecast for the rate of inflation from model (1) of Table 1 as a proxy for the nominal rate of interest.

The estimates for these transfer functions are presented in Table 4. In this table it is shown that in the case of Argentina when i is greater than one, both the degrees of the polynomials estimated and the error variance are higher than when i is equal to or less than one. I have no explanation for this except, as mentioned above, Dy_t is a very noisy series and as i becomes large it magnifies the noise of the series of "nominal income." The last transfer function reported in Table 4 includes a second-order autoregressive process for the error term which introduces an appreciable reduction in the error variance.

In the case of Brazil, when the real income elasticity is relatively large (1 or 1.5), the adjusted R^2's are low. The largest R^2 is obtained with $i = 0.5$ and with a second-order polynomial in the disturbance term.

It should be noted that if the estimate for the parameter a is assumed to be zero and if the real income elasticity of the demand for money is assumed to be one, our system reduces to a special formulation of Friedman's Theory of Nominal Income. This can be explained as follows: if a is assumed equal to zero, equation (1) can no longer be used to break down the changes in nominal income (obtained from equation (3)) between prices and output. So our system only explains nominal income.

In this case, equation (12) determines the price expectations (still under the hypothesis of rational expectations) that would dominate the changes in the nominal rate of interest in equation (3). Recall that from (12) we obtain the proxy Dr_t' for the nominal rate of interest.

IV. Conclusions

As indicated in the title of this paper, I have tried to explain the short-run dynamics of prices and output. An indicator of the degree to which this objective has been achieved could be the part of the variance in prices and output that has been explained by the model. In other words, we could look at the R^2's obtained in our transfer functions or regressions. For the case of Argentina, the R^2's for prices and nominal income have been close to .50, while for detrended output the R^2's have been in the order of .35. In the case of Brazil we obtained R^2's around .45 for prices, .30 for nominal income, and .43 for detrended real income.

[12]It is important to mention here an interesting result obtained in the work of Lucas (1973). He found, in a sample of 18 countries and working with annual observations, that "In a stable price country like the United States, then, policies which increase nominal income tend to have a large initial effect on real output, together with a small, positive initial effect on the rate of inflation. Thus the apparent short-term tradeoff is favorable, as long as it remains unused. In contrast, in a volatile price country like Argentina, nominal income changes are associated with equal, contemporaneous price movements with no discernible effect on real output" (see Lucas, pp. 332–33). Our results for Argentina and Brazil tend to confirm this finding and the underlying theory that specify that a favorable tradeoff between output and inflation depends upon "fooling" suppliers, a thing that becomes hard to do when the variance of the demand shifts becomes large.

TABLE 4—ESTIMATED TRANSFER FUNCTIONS FOR NOMINAL INCOME
(Estimates of Equation (3))

Model	Estimates of the AR and MA parts	Residual Sum of Squares (RSS); Degrees of Freedom (DF); RSS/DF; Adj. R^2	Estimates of Dr_t^i
	Argentina 1956-I-1973-II		
(1)	Dm_t: $(1.021 - 0.264L + 1.283L^2)/(1 + 1.739L + 0.596L^2)$	0.249863	0.565
(i = 1.5)	(0.269) (0.311) (0.247) (0.208) (0.234)	56	(0.311)
		0.00446	
		.27	
(2)	Dm_t: $(0.406 - 0.147L)/(1 - 1.496L + 0.737L^2)$	0.140220	0.504
(i = 1)	(0.149) (0.182) (0.120) (0.106)	58	(0.213)
		0.00242	
		.38	
(3)	Dm_t: $(0.424 - 0.130L)/(1 - 1.436L + 0.737L^2)$	0.139771	0.475
(i = 0.5)	(0.165) (0.212) (0.120) (0.106)	58	(0.215)
		0.00241	
		.38	
(4)	Dm_t: $(0.442 - 0.141L)/(1 - 1.407L + 0.722L^2)$	0.121832	0.266
(i = 0.5)	(0.194) (0.260) (0.176) (0.141)	54	(0.216)
		0.00225	
		.41	
	u_t: $1/(1 - 0.276L - 0.171L^2)$		
	(0.141) (0.150)		
	Brazil 1955-I-1971-IV		
(1)	Dm_t: $(0.212 + 0.103L + 0.348L^2)/(1 - 0.383L)$	0.377770	−0.093
(i = 1.5)	(0.348) (0.526) (0.521) (0.422)	57	(0.516)
		0.00593	
		.093	
(2)	Dm_t: $(0.253 - 0.272L + 0.356L^2)/(1 - 0.419L)$	0.182688	0.115
(i = 1)	(0.255) (0.389) (0.359) (0.306)	57	(0.380)
		0.00320	
		.171	
(3)	Dm_t: $(0.292 - 0.161L + 0.366L^2)/(1 - 0.466L)$	0.095822	0.723
(i = 0.5)	(0.184) (0.285) (0.243) (0.213)	57	(0.375)
		0.00168	
		.293	
(4)	Dm_t: $(0.197 - 0.061L + 0.221L^2)/(1 - 0.613L)$	0.086180	0.085
(i = 0.5)	(0.182) (0.272) (0.251) (0.226)	53	(0.350)
		0.00162	
		.342	
	u_t: $1/(1 - 0.223L - 0.169L^2)$		
	(0.167) (0.169)		

Notes: See Table 1.

Other indicators are the standard errors of the estimates and the t values. Standard errors have been reported in the tables of Section III. Not all the estimates of the parameters are significantly different from zero at the 0.05 level but many of them are indeed significantly different from zero at the 0.05 level in a two-tailed test. Other estimates are small in absolute value and not significantly different from zero—for example, in the case of the estimates of $Dy_{c,t-1}$ and c in the transfer functions for prices and expected prices. Such findings are not inconsistent with the theoretical model. As mentioned above, these parameters can be close to zero. Finally, there are other parameters that have large standard errors; in particular the slope coefficient of

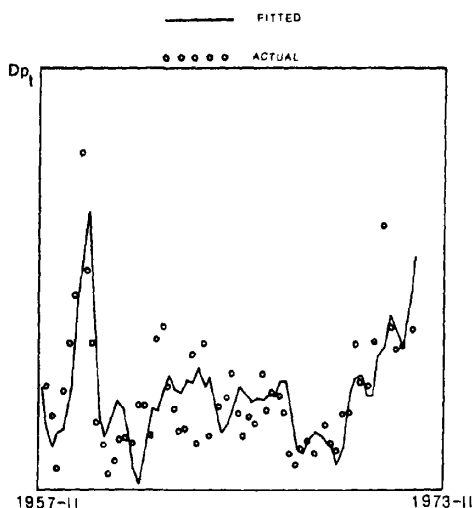



FIGURE 1. ARGENTINA: ACTUAL RATE AND FITTED VALUES FOR THE RATE OF CHANGE IN PRICES

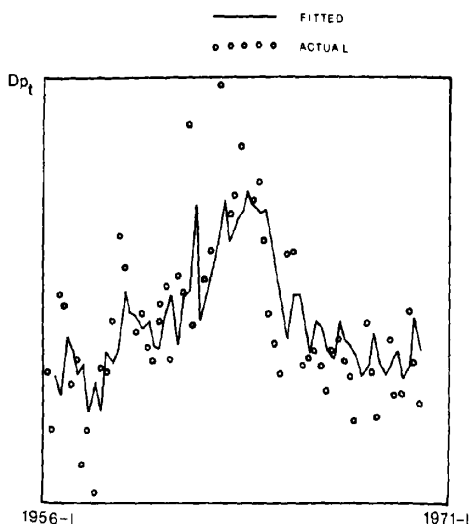



FIGURE 2. BRAZIL: ACTUAL AND FITTED VALUES FOR THE RATE OF CHANGE IN PRICES

our short-run Phillips curve, indicating that this relationship is empirically unstable.

In general the estimates for Argentina are more precise than the estimates obtained for Brazil. In both countries better fits were obtained for the rate of change in prices than for detrended income. The good performance of the model in explaining the rate of inflation can be illustrated by plotting the actual and fitted values from the reduced form for prices. This is shown in Figures 1 and 2. Figure 1 shows the case for Argentina. Here we observe that the model behaves well in explaining inflation, and only in two observations—one near the beginning and one near the end of the period—does the observed rate of inflation differ substantially from the fitted value.

Figure 2 illustrates the case of Brazil. Here we also observe the good performance of the model in explaining the large oscillations of the rate of inflation. Only in a few observations near the middle of the period do actual values differ substantially from the fitted values.

Although our results seem to be good relative to many other empirical studies working with highly noisy seasonally adjusted quarterly series, we still are not sure that we have really separated the true signal from the noise. That is, in explaining the movements of output apart from its long-run trend we have only used monetary shocks that impeded a correct anticipation of prices, and in this sense people were surprised (or fooled) during short periods of time.

From a theoretical point of view it can be said that our model makes use of two relatively new aspects of macroeconomic theory. One is the hypothesis of rational expectations; and the other is a sort of Phillips equation playing the role of the "missing equation" which, according to Friedman, states the difference between the quantity theory of money and the Keynesian income-expenditure theory.

REFERENCES

George E. P. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control,* San Francisco 1970.

R. B. Fernandez, "Short Run Dynamics of Output and Prices," unpublished doctoral dissertation, Univ. Chicago 1975.

Milton Friedman, *A Theoretical Framework for Monetary Analysis*, Occas. Paper 112, Nat. Bur. Econ. Res., New York 1971.

A. C. Goncalves, "The Problem of Stopping Inflation," unpublished doctoral dissertation, Univ. Chicago 1974.

R. E. Lucas, "Some International Evidence on Output Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *J. Soc. Ind. and Appl. Math.*, June 1963, *2*, 431–41.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

T. Sargent, "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment," *Brookings Papers*, Washington 1973, *2*, 429–80.

——— **and N. Wallace,** "Rational Expectations, the Optimal Monetary Instrument and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

A. Zellner, and F. Palm, "Time Series Analysis and Simultaneous Equation Models," *J. Econometrics*, May 1974, *2*, 17–54.

International Monetary Fund, *International Financial Statistics (IFS)* Washington, various years.

Instituto Nacional de Estadistica y Censos, *Boletin de Estadistica*, Argentina, various years.

Stigler, Kindahl, and Means on Administered Prices

By LEONARD W. WEISS*

Some years ago, George Stigler and James Kindahl (1970) published a remarkable set of transactions price series that they had compiled from purchasers' records. On the basis of those price data, they concluded that the phenomenon of "administered prices" was largely an illusion found only in the Bureau of Labor Statistics (*BLS*) wholesale price index data. Gardiner Means responded in an article in this *Review* (1972) that in fact the transactions prices compiled by Stigler and Kindahl supported the administered-price hypothesis. Stigler and Kindahl (1973) replied that Means had arbitrarily defined 15 of their price series as "market dominated," that he had redefined the business cycle turning points for purposes of his comparisons, and that he took as support for the administered-price hypothesis *any* deviation from classical predictions. This paper is an attempt to reconcile the two interpretations of the Stigler and Kindahl data.

I. The Usefulness of Price Statistics

The first issue considered is how representative the Stigler-Kindahl National Bureau (*NB*) price series are of actual transactions prices. The *NB* series are for large buyers often buying on long-term contracts. The *BLS* prices are spot list prices. Transactions probably occur at both prices. How nearly representative of prevailing prices are the two series?

A third set of prices which might serve to answer this question are the unit values computed by the Census in the 1963 Census of Manufacturers. These are available for 1958 and 1963 for many five-digit products. I was able to match 40 five-digit products to one or more of the 64 price series published by Stigler and Kindahl. Unit values represent indexes of values of shipments divided by indexes of physical output. They therefore amount to weighted averages of all transactions prices for the five-digit products involved where the weights are the amounts of individual transactions. The main problem with unit values is that they can change because of changes in the product mix contained in a five-digit product class as well as because of changing price levels. Moreover, many five-digit products contain other products besides those for which price series were available. While the usefulness of unit values for our purposes is open to some question, they do provide the one independent test of the two price series.

Simple averages of 1958 and 1963 monthly values for the *NB* and the *BLS* series were calculated and compared with 1963 unit values divided by 1958 unit values. The results appear in Table 1. All three indexes show declines between 1958 and 1963. The greater decline in the unit values is due mainly to one observation, "elemental gases," where the unit value fell by more than half while *BLS* and *NB* oxygen prices rose slightly. When that observation is excluded the means are similar. With that observation included, the *BLS* and *NB* series are more closely correlated with each other than either is correlated with the unit value index. With it excluded, the three correlations are similar.

All of the correlations are significant at a 1 percent level, implying that all three indexes contain similar information. The *NB* indexes are more closely correlated with the unit values than are the *BLS* indexes for the entire sample, but the *BLS* index fits a bit better when the outlier is excluded. The *BLS* and *NB* series both seem to contain elements of the pattern of prices shown in the unit values.

Most studies related to the administered-price hypothesis have correlated *BLS* price

*University of Wisconsin-Madison.

TABLE 1—MEANS AND CORRELATIONS COMPARING THE *NB* AND *BLS* INDEXES WITH UNIT VALUES 1958–63

	BLS	*NB*	Unit Values
For Forty Series:			
Mean	98.48	97.00	94.00
Simple Correlations			
BLS	1.000		
NB	.796	1.000	
Unit Value	.496	.524	1.000
Excluding Elemental Gases:			
Mean	98.31	96.90	95.23
Simple Correlations			
BLS	1.000		
NB	.797	1.000	
Unit Value	.744	.725	1.000

changes with concentration. The substantial differences between the *BLS* and *NB* indexes raise questions about the usefulness of those studies. Some comparisons made in Table 2 attempt to test the reliability of the *BLS* series. The first row of the table shows correlations between BLS_{t+1}/BLS_t and NB_{t+1}/NB_t for the two recessions and the two recoveries during the period covered by the Stigler and Kindahl data. Three different versions of the final recovery period are used for reasons explained below. Changes in the two price series are closely correlated in all six periods used.

A question that Stigler and Kindahl did not address is whether there is any bias in the relation between concentration and reported *BLS* and *NB* price changes. The second and third rows of Table 2 are meant to thrown light on that question. Here the ratio of BLS_{t+1}/BLS_t and NB_{t+1}/NB_t is related to the 1963 concentration ratio and to corrected 1963 concentration ratios.[1] In all but the first recovery period, the ratio had a negative relationship to concentration, meaning that *BLS* prices rose less or fell more with concentration than the *NB* prices did, but in no period was the correlation significant.

Altogether, the *BLS* and the *NB* price changes reflect each other reasonably well and their relationship to concentration is not importantly biased. Apparently, we can get something useful from working with *BLS* data.

II. Classifying the Data

The main concern of this paper is whether the *NB* series supports the administered-price hypothesis or not. An important issue in the debate between Means and Stigler and Kindahl is which of the prices covered in the study are "administered." Means listed 15 that he felt were not on the basis of judgment with respect to the character of the product. Stigler and Kindahl accepted only 3 "market-dominated" prices in their data, those for bituminous coal, plywood, and car flooring. Actually, both lists are arbitrary.

The natural criterion to apply is that used by Means in his previous work, the number of changes in the *BLS* price indexes. The number of price changes is not very interesting itself, but it seemed to correlate well with the extent of price change in Means' studies of the Great Depression. When Means first developed the concept of "administered prices" in 1935, he classified products into four groups: those that changed less than once every ten months; those that changed more than once in ten months but less than one time in four months; those that changed more than once in four months but less than three times in four months; and those that changed at least three times in four months (Means, 1935, p. 79). He didn't say at the time which series he considered to be "administered" prices and which not, but in later work referring to the same date, he seemed to count the first two classes as indicating administered prices, the last as "market-dominated" prices, and the third class as "intermediate" (Means, 1962, pp. 105–07). At any rate, he

[1] The corrections in the concentration ratios consist of 1) the combined concentration ratio wherever the product is primary to more than one industry and where the Census reports a combined concentration ratio, 2) a weighted average of regional four-digit concentration ratios for products that sell on regional markets and where the Census reports regional concentration ratios (paper boxes, paint, petroleum refining, cement, and bitumenous coal—data for coal is for 1964 and was supplied by the Federal Trade Commission (*FTC*); and 3) the Canadian concentration ratio for newsprint (from data supplied by the *FTC*).

TABLE 2—CORRELATION COEFFICIENTS RELATING *BLS* PRICE CHANGES TO *NB* PRICE CHANGES AND THE RATIOS OF THE TWO TO CONCENTRATION

Correlated Variables	1/57–4/58	4/58–5/60	5/60–2/61	2/61–3/62	11/64–11/66	2/61–12/66
$\frac{BLS_{t+1}}{BLS_t}$ and $\frac{NB_{t+1}}{NB_t}$	.774	.882	.781	.885	.776	.736
$\frac{BLS_{t+1}}{BLS_t} \Big/ \frac{NB_{t+1}}{NB_t}$ and the 1963 concentration ratio	−.207	.175	−.042	−.160	−.080	−.087
$\frac{BLS_{t+1}}{BLS_t} \Big/ \frac{NB_{t+1}}{NB_t}$ and the corrected concentration ratio	−.092	.129	−.148	−.155	−.177	.047
Number of Observations	62	62	62	62	64	64

had consistantly treated the administration of prices as a matter of degree, with administered prices shading into market-dominated prices. (Means, 1935, p. 82, and Means, 1939, p. 147.)

An adequate evaluation of the Stigler-Kindahl data requires that their price series be properly classified. Using the number of changes in the *BLS* prices during 1957–66 as a criterion, the price series reported by Stigler and Kindahl contain 3 market-dominated price series, 16 intermediate price series and 45 administered-price series. The market-dominated and intermediate price series are listed in Table 3 with the number of price changes reported in the

TABLE 3—MARKET-DOMINATED AND INTERMEDIATE PRICE SERIES INCLUDED IN STIGLER AND KINDAHL'S DATA

Product	Number of Price Changes 1957 1966	1963 Concentration Ratio	Corrected 1963 Concentration Ratio
Market-Dominated Price Series			
C-20 Regular Gasoline	98	33	51
C-21 Diesel and Distillate Fuel	95	36	51
C-63 Plywood	112	28	28
Intermediate Price Series			
C-10 Aluminum Ingot and Shot	36	76	76
C-13 Copper Ingot	42	84	51
C-14 Copper Pipe and Tubing	81	47	47
C-15 Copper Wire and Cable, Bare	38	50	50
C-16 Copper Insulated Wire	77	37	29
C-19 Brass Bars and Rod	56	47	47
C-22 Residual Fuel Oil No. 6	73	35	51
C-23 Bitumenous Coal	81	24	29
C-29 Book, Magazine Paper	33	29	29
C-31 Course Paper and Bags, Kraft Paper	45	38	38
C-32 Paper Board, Unfabricated	35	32	32
C-33 Paper Boxes and Shipping Containers	36	21	42
C-57 Paint	34	24	34
C-58 Cement	42	29	72
C-61 Electric Motors	62	47	47
C-64 Car Flooring	85	16	16

BLS series, their 1963 concentration ratios, and their 1963 corrected concentration ratios. There are 119 possible price changes in the ten-year period covered by the Stigler and Kindahl data.

All of the market-dominated price series are associated with products where concentration is low. The same is true of most of the intermediate price series, though aluminum and copper ingot and cement are certainly highly concentrated. Aluminum and copper are probably in this category because of competition from imports and scrap, but cement has no such explanation. The average values for concentration are 36.1 for the 19 products in Table 3 and 58.6 for the remaining products in the Stigler-Kindahl sample. The average values for the corrected concentration ratios are 40.7 for the products in Table 3 and 57.2 for the rest of the sample. For the most part, the market-dominated and intermediate products have concentration characteristics that most economists would expect.

Whether the intermediate price series should be treated as administered or not is a matter of judgment. In his previous studies, Means treated administered prices as a matter of degree, examining separately the changes that occurred in price series where price changes had been of various frequencies. The same will be done here.

Means chose somewhat different periods from those used by Stigler and Kindahl in evaluating the direction of *NB* price change. The two studies agreed on the dates of the two contractions but they chose different recovery periods. Means ended the recovery from the 1958 recession six months before the start of the 1960 recession. Stigler and Kindahl's adoption of the regular reference cycle seems less arbitrary.

The really serious problem arises with the recovery from the 1960 recession. That expansion continues for nine years, a period so long that both Means and Stigler-Kindahl chose shorter subperiods. Means used the immediate recovery period from February 1961 through February 1962. Stigler and Kindahl chose November 1964 through November 1966. We will use both periods and the full recovery period as well.

III. Theoretical Expectations

Just what to expect during recessions and recoveries is also a matter of controversy. There are several versions of the administered-price hypothesis. Means seemed to picture any pattern of price change that did not follow classical expectations as confirming the administered-price hypothesis. This left him taking declining prices in a period of expansion as supporting it. Stigler and Kindahl took the other extreme, requiring that prices remain fixed in the face of short-run fluctuations if the hypothesis is to be confirmed. Realistically, Means' studies in the 1930's (1935; 1939) pictured administered prices as changing relatively less than market-dominated prices during business fluctuations. This last will be the hypothesis to be tested here.

The implications of the administered-price hypothesis are least equivocal in recessions. In such periods the hypothesis would be confirmed if prices remained stable or rose. A more sophisticated version would require that we control for changes in direct costs, but this cannot be tested here for lack of data on direct costs for five-digit products during the periods covered. As an alternative, we will examine deviations from trends. The trends in price should reflect cost changes.

The prediction for a period of expansion is more equivocal. Presumably, market-dominated prices would rise during such periods, but what should we expect of administered prices? A failure of price to change would be evidence of administered prices, but surely one would not expect a *decrease* in price under such circumstances. Perhaps the safest expectation is that administered prices would rise less than market-dominated prices in a recovery period immediately following a recession in which they fell less. This might argue for the earlier recovery period chosen by Means, from February 1961 through February

1962, over the considerably later recovery period chosen by Stigler and Kindahl (November 1964–November 1966), but the prediction for any recovery period is much less certain than for a contraction.

Stigler and Kindahl express doubt about the theoretical basis for the predictions of the administered-price hypothesis:

> Classical theory leads one to expect prices to fall in competitive industries during a business contraction because both demand and marginal production costs fall, and that reverse movements will occur in expansions. This expectation was not subjected to elaborate analysis perhaps because a similar pattern was expected in monopoly. Here too, marginal costs would fall and there is no strong reason to expect marginal revenue to rise, although price reduction was no longer a *necessary* result of a leftward shift in demand and cost functions. [1970, pp. 60–61]

It really isn't very hard to find a theoretical basis for relatively rigid monopoly price in the presence of contraction, however. A monopolist must make explicit decisions about price. It is not given to him by the market. He will presumably make prices in a way that is expected to maximize the present value of the firm. This means he will price with an eye to possibilities for substitutes, potential entry, and similar long-run considerations. A short-run fall in demand as might occur in a recession need not lead to a downward shift in price. The short-run elasticity of demand is surely less than the long-run elasticity, and in many cases must be less than one. This would imply that a monopolist considering a price cut to deal with a temporary recession would face a lower (and possibly negative) marginal revenue compared to the marginal revenue he would face in making general price policy. The natural conclusion seems to be that optimal price policy in a recession would often be to hold prices stable or to change them only in response to changes in long-run expectations. It would follow that price increases in recovery periods would not be profitable unless they also reflected long-run expectations.

Stable rather than declining prices during a recession would also probably be more profitable for the members of a competitive industry taken as a whole, but of course no individual firm can prevent the price decline that would naturally occur in such a market. Given such a decline during a recession, one would surely expect a relatively large increase in a subsequent recovery in competitive markets. Means himself felt that there was no necessary relationship between administered prices and monopoly, but his example of a competitive industry with administered prices was automobiles—obviously an oligopolistic industry. In my opinion, some market power is necessary for administered prices. Prices should be more rigid the greater the level of concentration.

IV. The Evidence on Administered Prices

Table 4 shows the directions of change and of deviations from trends for *NB* prices during the 1958 and 1960 recessions for the three categories of price series. In general, the bulk of the market-dominated and intermediate series declined during the two recessions. By contrast, a majority of the administered prices either rose or remained constant (changed by less than .05 percent per month) during the 1958 recession. The contrast between the administered prices and the others is even greater when the deviations from trend are examined. The experience during the 1958 recession strongly supports the administered-price hypothesis.

The evidence for the 1960 recession is less one-sided. Two of the three market-dominated prices rose and more of the administered prices fell, both in absolute terms or relation to trend. Still, 43 percent of the administered prices rose or remained unchanged compared with 26 percent of the market-dominated and intermediate prices taken together, and the average decline in the administered prices was less than in the intermediate prices. Altogether, the results for the two recessions give considerable support for the administered-price hypothesis.

TABLE 4—DIRECTION OF CHANGES AND AVERAGE CHANGE IN *NB* PRICES DURING THE TWO RECESSIONS

	Changes in Raw Indexes		Deviations From Trend	
	7/57–4/58	5/60–2/61	7/57–4/58	5/60–2/61
Market-Dominated Prices				
Increase	0	2	0	2
No Change	0	0	0	0
Decline	3	1	3	1
Intermediate Prices				
Increase	1	2	2	2
No Change	1	1	1	0
Decline	14	13	13	14
Administered Prices				
Increase	12	7	21	12
No Change	22	12	9	7
Decline	10	25	13	24
Average Changes[a]				
Market-Dominated Prices	−3.5	+0.2	−2.8	+1.0
Intermediate Prices	−6.1	−3.6	−6.8	−3.8
Administered Prices	−0.1	−2.2	−1.1	−0.8

[a] Shown in percent.

The results for the recovery periods are shown in Table 5. During the 1958–60 recovery, two of the three market-dominated prices fell but a larger proportion of the intermediate prices rose than did the administered prices, and the average increase for intermediate prices was larger than for the administered prices. Deviations from trend were preponderantly positive for both intermediate and administered prices.

TABLE 5—DIRECTION AND EXTENT OF *NB* PRICE CHANGES AND DEVIATIONS FROM TRENDS DURING RECOVERY PERIODS

	Changes in Raw Indexes				Deviations from Trends			
	4/58–5/60	2/61–3/62	11/64–11/66	2/61–12/66	4/58–5/60	2/61–3/62	11/64–11/66	2/61–12/66
Market-Dominated Prices								
Increase	1	0	3	0	1	1	3	2
No Change	0	1	0	2	1	0	0	1
Decrease	2	2	0	1	1	2	0	0
Intermediate Prices								
Increase	11	6	13	9	11	7	13	11
No Change	1	5	2	5	4	3	3	3
Decrease	4	5	1	2	1	6	0	2
Administered Prices								
Increase	17	4	21	9	31	10	25	16
No Change	9	13	9	8	6	10	11	24
Decrease	18	27	15	27	6	23	8	4
Average Change in Price[a]								
Market-Dominated Prices	−1.7	−1.7	+2.2	−3.8	+0.5	−1.4	+4.3	+2.4
Intermediate Prices	+7.8	+0.4	+11.6	+13.6	+7.0	−0.1	+10.8	+11.0
Administered Prices	+1.7	−1.9	+4.2	−2.8	+3.1	−1.8	+4.9	+2.9

[a] Shown in percent.

The experience for the long recovery from the 1960 recession differs with the portion of that recovery analyzed. In the initial recovery period (1961–62) emphasized by Means, a majority of the administered prices declined both absolutely and relative to trend which is certainly not what the classical analysis would suggest. It really is not what the administered-price hypothesis would suggest either. The three market-dominated prices fell, on average, and the intermediate prices showed no net tendency to either rise or fall.

In the 1964–66 period emphasized by Stigler and Kindahl, a larger number of administered prices increased, and a majority had positive deviations from their trends. However, the market-dominated and intermediate prices conformed more closely to the classical hypothesis.

Over the long period of recovery from 1961 to 1966, the majority of administered prices fell as did one market-dominated price and two of the intermediate prices. On average, both market-dominated and administered prices fell, though intermediate prices rose substantially during the long recovery. As might be expected for so long a period, almost half the series deviated from trend by less than 0.05 percent per month, but the intermediate prices rose above trend much more than the administered prices did.

The long period is probably not the right one for studying administered prices. Oligopoly prices can be expected to change over the long run in response to changes in costs or in the elasticity of demand. The period from 1961 to 1966 must be pretty close to the long run. The administered-price hypothesis is best tested in a shorter period.

Altogether, the evidence during the recovery periods is more equivocal than that during the recessions. The intermediate price series comes closer to fitting the classical hypothesis than the other two series. The behavior of the administered-price series certainly doesn't fit the classical hypothesis. On balance, the evidence for the recovery periods comes closer to fitting Means' conclusions than those of Stigler and Kindahl. Yet the results are basically equivocal. Neither hypothesis predicts price decrease in time of recovery, and there were many price decreases among administered prices.

V. Concentration and Price Change

Concentration does not play a direct role in the administered-price hypothesis as Means expressed it. In his original discussion of the topic in 1935, he said:

> Administered prices should not be confused with monopoly. The presence of administered prices does not indicate the presence of monopoly, nor do market prices indicate the absence of monopoly. In many highly competitive industries, such as the automobile industry, prices are made administratively and held for fairly long periods of time. On the other hand, it is conceivable that in a monopolized industry, the product might be turned out according to some fixed production schedule and sold for what it would bring on the market regardless of price.

TABLE 6– RECESSION PERIOD CHANGES IN PRICES AND DEVIATIONS FROM TRENDS CLASSIFIED BY DEGREE OF CONCENTRATION

	Changes in Raw Indexes		Deviations From Trends	
	4/58–5/60	2/61–3/62	4/58–5/60	2/61–3/62
$CR \leq 30$				
Increase	0	1	0	1
No Change	0	0	1	0
Decrease	6	5	5	5
$30 < CR \leq 50$				
Increase	5	5	9	8
No Change	6	4	1	2
Decrease	11	13	11	11
$CR > 50$				
Increase	6	6	14	8
No Change	17	8	7	5
Decrease	12	21	14	22
Average Change in Price[a]				
$CR \leq 30$	–.023	–.020	–.023	–.020
$30 < CR \leq 50$	–.019	–.022	–.001	–.010
$CR > 50$	–.015	–.023	–.007	–.017

[a]Shown in percent.

TABLE 7—DIRECTION AND EXTENT OF *NB* PRICE CHANGES AND DEVIATIONS FROM TREND DURING RECOVERY PERIODS CLASSIFIED BY CONCENTRATION

	Change in Price				Deviations from Trend			
	4/58–5/60	2/61–3/62	11/64–11/66	2/61–12/66	4/58–5/60	2/61–3/62	11/64 11/66	2/61–12/66
$CR \leq 30$								
Increase	5	2	6	3	5	2	6	4
No Change	0	1	0	3	1	2	0	1
Decrease	1	3	0	0	0	2	0	1
$30 < CR \leq 50$								
Increase	8	4	14	7	12	7	13	9
No Change	2	7	3	6	4	4	5	9
Decrease	12	11	5	9	5	10	3	3
$CR > 50$								
Increase	17	4	18	9	26	9	21	13
No Change	7	11	6	6	6	8	9	19
Decrease	11	20	12	21	3	18	6	4
Average Change in Price[a]								
$CR \leq 30$	4.3	0.2	5 9	5.2	4.5	0.3	6.0	5.6
$30 < CR \leq 50$	−0.3	−3.4	5.5	−1.6	2.5	1.3	8.1	5.4
$CR > 50$	2.4	−2.6	3.1	−3 2	4.8	1.2	5.5	3.7

[a]Shown in percent.

> In general, monopolized industries have administered prices, but so also do a great number of vigorously competitive industries in which the number of competitors are few. [pp. 78–79]

Nevertheless, most who have studied the subject since Means have looked for a relationship between concentration and price behavior. Administered prices are clearly inconsistant with pure competition. At least some market power is needed to administer a price.

Table 6 analyzes changes in prices and deviations from trends during the two recession periods with the price series classified by concentration. Products with concentration ratios of 30 or less clearly sell on unconcentrated markets; those with concentration ratios above 50 sell on concentrated markets; and those in between are classified as mixed.

Prices in unconcentrated markets were more inclined to fall both absolutely and relative to trend than other price series, but there were only six unconcentrated markets represented. No distinction can be drawn between the concentrated and the mixed price series.

Table 7 shows a similar analysis for the recovery periods. No distinctions among the three categories seem warranted in these periods.[2]

Table 8 shows correlations of price changes with concentration and corrected concentration. Neither concentration nor corrected concentration is significantly related to price change in any of the periods covered. This is not a very serious blow for the administered-price hypothesis, however. The studies done previously with *BLS* series

TABLE 8—CORRELATIONS BETWEEN *NB* PRICE CHANGES AND CONCENTRATION IN SIX PERIODS

Period	Concentration Ratio	Corrected Concentration Ratio
7/57–4/58	.089	.198
4/58–5/60	.149	.034
5/60–2/61	−.156	−.037
2/61–3/62	−.012	−.078
11/64–11/66	−.015	−.098
2/61–12/66	−.132	−.130

[2]This is apparently not due to inaccuracies in the measurement of concentration. When the price series were classified by corrected concentration ratios, the results were no less equivocal.

show only a weak effect for concentration if changes in direct costs are not included as well. (See Horace DePodwin and Richard Selden; the author.) Moreover, those studies generally show no significant effect for concentration over the period 1959–1966. (See the author; R. W. McLaren; J. Fred Weston and Steven Lustgarten, 1974).

VI. The Frequency of Price Changes

The hallmark of administered prices according to Means is large infrequent jumps in price with long periods of rigid prices in between. This appears in many *BLS* series, especially those identified as administered prices. The *NB* series prepared by Stigler and Kindahl do not exhibit such a pattern except for steel products. This need not mean that the underlying price series from which the *NB* series are constructed are gradually changing prices. Indeed, as Means points out (1972, p. 301), the one group of underlying series that Stigler and Kindahl present (for ammonia) do show long periods of rigid prices followed by sharp jumps. One series does not change at all in eight years. The other three series show only 3, 5, and 15 changes, and 7 of the 15 changes in the last are uniform seasonal changes. Each of the four series presented would by itself fall into the administered price category because it shows less than 30 changes over the ten-year period. To some extent the long periods of stability in these prices indicate long-term contracts which reflect both buyer and seller expectations. Stigler and Kindahl do not indicate where this is the case, so the analysis of other series is necessarily uncertain.

The small and continuous changes in the *NB* series as contrasted with the *BLS* series and at least the underlying ammonia series reported by Stigler and Kindahl could easily be attributable to the averaging of price changes in many contracts with different termination dates, and the interpolations of series with breaks in them.

The frequency of change is not an important feature of a price in and of itself. It is a characteristic that Means used to classify prices as more or less administered and which seemed to predict quite well which type of price fell most during the Great Depression. It also seems to have predicted which prices fell most during the two recessions covered by the Stigler-Kindahl study whether *BLS* or *NB* prices are used. The failure of their published data to reveal many differences in frequency of price change does not make the frequency classification based on *BLS* data less useful.

VII. The Unique Price

What the Stigler-Kindahl data (reported for ammonia and summarized for other series) do reveal is that the commonly held belief in a unique price, at least for undifferentiated products with expert buyers, is unjustified. This was certainly the type of product that they examined. They report that precisely that type of product virtually never shows a unique price. Of the 64 commodities for which they publish data, they give standard deviations of price changes of individual buyer series for 633 commodity-years.[3] Of these, only 8 commodity-years have standard deviations of zero as would be expected if a unique price prevailed. Thirty-three commodity-years showed standard deviations of less than 0.1. These were heavily concentrated in steel where price discipline seems to be strong, but even there standard deviations exceeded 1.0 in 20 percent of the years. For the remaining 543 commodity-years the standard deviations exceeded 1.0 in 49.3 percent of the cases.

One of the most intriguing of their results is that the dispersion of individual buyer series price changes increases with the variability of the *NB* index over time and with the number of reporters, and decreases with the degree of concentration for the product involved. All three tendencies were highly

[3]Ten years for each commodity except electric batteries where data was available for only five years, and copper ingot where there was only 1 firm reported in two years.

significant.[4] Stigler and Kindahl interpret the negative effect of concentration as due to the increasing difficulty of policing price agreements as concentration falls (1970, p. 89).

VIII. Conclusion

The conclusion of this study would seem to be as follows: The *NB* and *BLS* series do not differ importantly. The *NB* is about as closely correlated with census unit values as the *BLS* series, but all three are closely correlated, implying that all three contain meaningful information. There is no significant bias in the two sets of series with respect to concentration so that studies using *BLS* series are meaningful.

More important, the *NB* series do follow patterns that support the administered-price hypothesis. For those series where the *BLS* indexes show few changes, the *NB* series did increase or remain constant in a majority of observations during the two recessions. Their performance during recovery periods is more ambiguous, but certainly does not correspond to the classical pattern. There does seem to be such a thing as an administered price. On the other hand, there was little or no relationship between concentration and changes in the *NB* series.

The *NB* series are a unique and valuable set of data that has been little used. There are more questions to be answered. What does account for deviations from list price? A sensible study might calculate year-to-year relative price changes in the *BLS* and *NB* series, correlate these over the ten-year period for individual series, and regress the correlation coefficients on structural characteristics that might explain price competitiveness such as concentration and variability of demand. Again, what accounts for differences in relative stability of transactions prices? One might study this by correlating variability of individual *NB* price series around trends with variables that might be expected to explain price stability such as concentration, demand variability relative to trend, and possibly distance from final demand. These questions will be left to others.

[4]Their regression equation was.

$$\sigma_{\Delta P} = \underset{(.203)}{.620} + \underset{(.073)}{.432}\,\sigma_P - \underset{(.0028)}{.0062}C + \underset{(.0041)}{.018}N$$

where $\sigma_{\Delta p}$ is the standard deviation of price changes among individual series for a product during the first six months of 1965, σ_P is the intertermporal standard deviation of the *NB* series for the same period, C is concentration, and N is the number of buyer reporters (Stigler and Kindahl, 1970, p. 91).

REFERENCES

H. DePodwin and R. Selden, "Business Pricing Policies and Inflation," *J. Polit. Econ.*, Apr. 1963, *71*, 116–27.

R. W. McLaren, Testimony, in Joint Economic Comm. hearings on "The 1970 Midyear Review of the State of the Economy," Part 1, 108–28, Washington 1970.

George C. Means, *Industrial Prices and Their Relative Inflexibility*, U.S. Senate Document 13, 74th Congress, 1st Session, Washington 1935.

———, *The Structure of the American Economy, Part I, Basic Characteristics*, National Resources Comm., June 1939.

———, *The Corporate Revolution in America*, New York 1962.

———, "The Administered-Price Thesis Confirmed," *Amer Econ Rev*, June 1972, *62*, 292–306.

George J. Stigler and James K. Kindahl, *The Behavior of Industrial Prices*, New York 1970.

——— **and** ———, "Industrial Prices, as Administered by Dr. Means," *Amer. Econ. Rev.*, Sept. 1973, *63*, 717–21.

L. W. Weiss, "Business Pricing Policies and Inflation Reconsidered," *J Polit. Econ.*, Apr. 1966, *74*, 177–87.

J. F. Weston and S. H. Lustgarten, "Concentration and Wage-Price Changes," in H. Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston 1974, 307–32.

Income and Urban Residence: An Analysis of Consumer Demand for Location

By WILLIAM C. WHEATON*

It is by now well documented that in American urban areas the vast majority of middle and upper income households live further from the city center in separate suburban communities (see Anthony Catenese; Edgar Hoover and Raymond Vernon). Despite popular impressions, this trend is not a recent one. It has been evolving gradually ever since the streetcar suburbs of the late nineteenth century (see Samuel Warner).

The consequences of this spatial pattern have been quite serious. The outward mobility of those with means has left American cities as segregated domains for the poor. Within city boundaries the poor can tax only themselves for necessary but deteriorating services (see Bennett Harrison). Having escaped this tax burden, middle income Americans enjoy a substantial "fiscal surplus" in the suburbs—providing an implicit and regressive redistribution of income. The spatial flight of the middle class has raised the issue of efficiency as well. The extensive deterioration and abandonment of inner city housing capital strikes many as being socially, if not privately, inefficient (see Jerome Rothenberg).

I

In searching for an explanation to these problems, many economists suggest that cause and effect are simultaneous. Urban fiscal disparities and city decay are not only the product of middle class flight, but the cause as well (see David Bradford and Harry Kelejian). The fragmented structure of American local government and the presence of racial or social externalities are the primary forces which have generated the spatial pattern to begin with (see Charles Tiebout; Robert Haugen and A. James Heins).

In contrast to these views, Richard Muth, William Alonso and others (Martin Beckmann, Edwin Mills) have developed a different argument to explain the increase in income with greater distance to the city. Using a simple model of spatial equilibrium, these authors argue that as consumers move farther from their center of employment, greater commuting costs must be counterbalanced by reduced expenditures on land. Wealthier consumers need reduce their offered land price by less because they spread such cost increases over more land As a consequence, they are willing to bid more for peripheral sites than the poor, at least relative to their offers for central locations. Since land must be allotted to the highest bidder, an equilibrium location pattern requires that income increase with greater distance from work.

As both Muth and Alonso have pointed out, this conclusion assumes that commuting "costs" (including time) are constant with respect to income. If the marginal cost of commuting increases with income, then wealthier consumers must reduce their land expenditure by a greater amount than poorer households. If the rate at which marginal travel costs increase with income is sufficiently large, then this can offset the aforementioned tendency for wealthier consumers to bid relatively more for further locations. In short, the net outcome of the argument really depends on how rapidly commuting outlays change with income relative to land consumption. With an income inelastic land demand and noticeably greater commuting costs for the wealthy, greater income leads to more central locations. If land demands are income elastic and commuting expenses relatively fixed,

*Assistant professor, departments of economics and urban studies and planning, Massachusetts Institute of Technology.

income should increase with distance from work.

This theory is quite attractive, for it offers an explanation of the spatial income gradient in other countries besides America. In European and Latin American cities, the poor usually inhabit the peripheral areas, while the rich and middle class live centrally (see Alonso). The presumption is that while American land demands are income elastic and the value of commuting time income inelastic, just the opposite pattern of preferences exists in these other societies. To date there has been no attempt to determine empirically whether these preference patterns exist. Do American demands for land really increase rapidly with income while commuting costs do not? The objective of this paper is to estimate empirically the Alonso-Muth spatial income effect and to determine whether this long-run equilibrium theory is a significant determinant of American land use patterns.

In the next section, the original arguments of Muth and Alonso are reviewed and updated. Section III discusses the estimation of land demand and travel costs for different income groups. In Section IV, these estimates are used to simulate a competitive locational equilibrium between income groups on the basis of preferences for land and travel. The results suggest that although higher income may tend to generate suburban locations, the differences in land bids that lead to this pattern are small and statistically insignificant. The long-run theory of a competitive land market is therefore at least a lesser determinant of the income-location pattern in America. The current spatial income gradient in metropolitan areas would seem to be more the result of social or racial externalities and the incentives produced by municipal decentralization.

II

In Muth's original discussion of income and location, consumers faced a known rent gradient, and by maximizing utility, selected that distance where the marginal savings in land expenditure exactly balanced the increase in commuting costs. Muth argued that if a consumer with greater income had the same marginal cost of travel, but higher land consumption, then the distance at which this first-order condition held would be greater. If, on the other hand, the wealthier consumer's travel costs were substantially higher while his land consumption was almost the same, then optimality would dictate a closer more central location.

Alonso's version of this argument was based on his so-called "bid price" approach to the urban land market. Viewing the market as an auction place, consumers offer bids for the use or purchase of land. Since landlords are inherently spatial monopolists, and consumers atomistic, each parcel of land goes to the highest bidder. The equilibrium rent profile is thus the outer envelope of such bidding. Since the bid price approach is used in this study, it is developed more formally below.

Consumer "bids" are more precisely defined as the maximum surplus per unit of land (rent) that can be extracted from households, subject to the constraint that such payments leave the household no worse off than others of similar means. This uniform level of welfare is denoted as u^o, and is obtained from the consumption of land q, other goods x, housing H, and the disutility of commuting distance t.

$$u^o = u(x, q, t, H) \tag{1}$$

The consumer's bid price for land is derived from his income y and his budget constraint (2). The bid price for land is R, while $c(H)$ is the cost of housing, and $k(t)$ is the money cost (as opposed to disutility) of traveling distance t.

$$R = \frac{y - k(t) - x - c(H)}{q} \tag{2}$$

With u^o and t taken as parameters, a consumer's bid price for land is the maximum value of (2) subject to (1). Assuming long-run flexibility, x, q, and H are all considered

variable. Letting λ be the Lagrangian associated with the constraint (1), bid price maximization yields the constraint (2) and the marginal conditions (3) below.

$$(3) \qquad \lambda \frac{\partial u}{\partial q} = R/q$$
$$\lambda \frac{\partial u}{\partial x} = 1/q$$
$$\lambda \frac{\partial u}{\partial H} = \frac{\partial c}{\partial H}/q$$

From the envelope theorem and the conditions above, the influence of distance on bid prices can be determined as:

$$(4) \quad \frac{dR}{dt} = \frac{\partial R}{\partial t} = \frac{-dk}{dt}/q + \lambda \frac{\partial u}{\partial t} = 1/q \left[\left(\frac{\partial u}{\partial t} \bigg/ \frac{\partial u}{\partial x} \right) - \frac{dk}{dt} \right] < 0$$

If land in the long run is allocated to those bidding the highest price for it, households with relatively flat bid price gradients must ultimately wind up locating at greater distances than households with bid price gradients that steeply decline over distance. This is not only an equilibrium, but an efficient arrangement as well, for a steeper bid price gradient means an individual places a higher marginal value on central locations.

To Alonso, then, the influence of income on location reduces to a question of how income y alters the slope of the bid price gradient (4). This, of course, depends on the same tradeoff discussed by Muth. On the one hand, greater income increases land consumption q, which from (4) is seen to make the bid price gradient flatter over distance. At the same time it will usually increase the marginal expenditure or income value of travel $(\partial u/\partial t/\partial u/\partial x)$, and this makes the offered price gradient steeper. If the total marginal cost of travel, the bracketed term in (4), is denoted by MCT, then at a fixed location the derivative of the bid price slope with respect to income is:

$$(5) \quad \frac{d^2R}{dtdy} = \frac{MCT}{yq}\left[-\frac{dq}{dy}\frac{y}{q} + \frac{dMCT}{dy}\frac{y}{MCT} \right]$$

If land is more income elastic than total marginal travel costs (MCT), then greater income reduces the slope of the bid price gradient, suggesting that the wealthy will live further from the center. If total marginal travel costs are more elastic, then bids become steeper with income, and the rich locate centrally.

Alonso acknowledged that bid price gradients must really be compared only in a general equilibrium solution. This necessitates determining an appropriate utility level (u^o in (1)) for each group of similar consumers. Again from the envelope theorem and condition (3):

$$(6) \qquad dR/du^o = -\lambda < 0$$

In order to bid more for land, a group of households must accept a lower level of welfare. This allows that group to capture or bid away an increasing amount of land from other groups. Since lower utility levels yield higher bid prices, they will also reduce the chosen levels of land consumption. The equilibrium solution is thus a set of utility levels, such that the area over which a household group has the highest bid exactly equals their aggregate desired land consumption.

In both Alonso's model and this empirical study the urban plain is featureless, all employment centrally located, and transportation continuous. Land development is thus circular and each household group (i) will locate its N_i members in one or more radial rings. Let $R_i(u_i, t)$ and $q_i(u_i, t)$ stand respectively for the bid price and land consumption gradients that are solutions to (2) and (3). Equilibrium requires determining a set of utility levels so that the areas commanded by each group, defined as ϵ_i in (7), just meet their land demands as defined in (8).

$$(7) \quad \epsilon_i = [t: R_i(u_i, t) > R_j(u_j, t) \qquad j \neq i]$$

$$(8) \quad 2\pi \int_{\epsilon_i} t/q_i(u_i, t)\,dt = N_i$$
over all groups i

The final resolution of the Muth-Alonso argument thus rests on being able to esti-

mate bid price and land consumption functions for different economic groups and then using these in a small simulation of competitive bidding. This will determine the locational order by distance that obtains in a general equilibrium, and consequently the relative steepness of different bid price gradients.

III

The solution to the long-run competitive bidding model described in the previous section depends on several pieces of information, but most importantly on the structural form and parameters of consumer utility functions. The standard way of estimating these is with a set of demand equations, and several recent studies have attempted this for housing and locational choice (see Gregory Ingram et al.; Mahlon Strazheim). A complementary technique developed by this author in a previous paper is based on a modification of the bid rent approach. This procedure recognizes that because of historical evolution, most cities do not achieve the kind of complete supply-demand equilibrium that characterizes the Alonso-Muth models. Since the housing stock is replaced or altered only gradually, its characteristics are essentially fixed in the short run. Given high mobility during this period, it is reasonable to expect that consumers will "trade" among one another to obtain an "exchange equilibrium." Without complete supply adjustment, ostensibly identical households will have to occupy different housing and pay appropriately different rents. Assuming equilibrium, the combination of rent and housing attributes that identical households consume must lie along an indifference surface. With sufficient data it should be possible to estimate consumer utility parameters by fitting this surface.

In order to derive the estimating equation implied by this approach, a group of identical households are hypothesized to have a utility function of the generalized CES form (9). The variables q and x are defined as in Section II, while T is the time spent in traveling distance t and the h_i are housing attributes of the vector H. This function has quite broad properties and includes the Cobb-Douglas form as a limiting case.

$$(9)\quad u^o = -x^{-\alpha_o} - \beta_T T(t)^{-\alpha_T} - \beta_q q^{-\alpha_q} - \sum_i \beta_i h_i^{-\alpha_i}$$

The consumer's budget constraint (10) can be incorporated into this utility function to yield the total offered rent (R) for any fixed combination of locational and housing attributes that exists in the short run (11). Naturally, such payments will depend on those attributes, transportation costs $k(t)$, and consumer income y.

$$(10)\quad y = R + k(t) + x$$

$$(11)\quad R = y - k(t) - \left[-u^o - \beta_T T(t)^{-\alpha_T} - \beta_q q^{-\alpha_q} - \sum_i \beta_i h_i^{-\alpha_i}\right]^{-1/\alpha_o}$$

For a sample of households which are defined to have identical income and tastes, equation (11) should hold across their different patterns of location and housing consumption. Assuming some stochastic specification, it should be possible to find the maximum likelihood estimates of the utility parameters α_T, β_T, α_q, β_q, α_i, β_i, α_o, and the utility level u^o.[1]

A critical issue in the development of this estimating procedure concerns the problem of how to group households so as to best meet the assumption of "identicalness." Clearly, various socioeconomic groups can be expected to have different preferences, so household samples should be homogeneous at least with respect to age or life cycle, household size, race, and occupation. More problematic, however, is the question of whether the sample should also be stratified by household income. If one postulates

[1] It is important to make clear that this estimating process does not assume cardinality of utility functions. Any order preserving transformation of (9), say $g(-)$, leaves (11) unchanged. The constant utility level is simply reinterpreted as $g^{-1}(u^o)$.

TABLE 1—ESTIMATED HOUSEHOLD PREFERENCES FOR LAND AND TRAVEL

Household Strata[a]	1	2	3	4	5	6	7
Income (thousands)	5-10	10-15	15-25	25+	5-10	5-10	10-15
Size	3-4	3-4	3-4	3-4	1-3	1-2	5-6
Age	30-55	30-55	30-55	30-55	-30	56+	30-55
Sample	106	144	100	45	120	79	69
R^2	.77	.67	.82	.52	.90	.62	.73
$\frac{\partial u}{\partial T} \Big/ \frac{\partial u}{\partial x}$: T = .5	14	73	166	512	103	42	79
T = 1.5	60	102	235	724	169	84	110
$\frac{\partial u}{\partial q} \Big/ \frac{\partial u}{\partial x}$, $q = .05$	6620	6390	12310	13150	2410	8702	50800
$q = .30$	1032	1048	1520	2148	401	1368	1192
q chosen at. R = \$20,000 T = .5	.163	.168	.246	.35	.066	21	.233
$\frac{dR}{dt} = \left[\left(\frac{dT}{dt}\frac{\partial u}{\partial T}\Big/\frac{\partial u}{\partial x}\right) - \frac{dk}{dt}\right]\Big/q$	-223	-247	-198	-215	-665	-187	-181
$\frac{dR}{dt} = \left[25y\frac{dT}{dt} - \frac{dk}{dt}\right]\Big/q$	-400	-410	-394	-408		–	–

[a] All are white, white-collar occupations

Note: T = hours of travel time each work day; x = dollars of "other" expenditure per year evaluated by strata at 5000, 8500, 16000, 39000, 5700, 5000, 8900, q = acres of land; dT/dt = .04 (25 mph); dk/dt = \$36/year (6¢/mile)

the existence of classical "income effects," it makes little sense to stratify groups by income as well as socioeconomic characteristics, for then the stratification becomes the explanation for the income effect. This suggests that equation (11) be estimated within socioeconomic groups, but across income levels. Of course, in this case u^o will no longer be constant, instead it will vary monotonically with income. This added relationship can be estimated within equation (11) and in principal presents no problem.[2]

The argument against this approach is that while income effects exist over time for particular individuals, in a cross-sectional sample the impact of different income may also reflect different backgrounds. Families of greater income usually have more education, acquired wealth, and different reference groups. As a consequence, there may be little reason to expect that preferences or utility parameters are independent of income in a cross-sectional sample. Instead of estimating a single utility function across income levels and then varying income to generate different bid rents, it seems less restrictive to estimate a separate utility function for each income group.

Using this estimating procedure for different socioeconomic strata seems ideally suited to the purposes of this study. In the short run, when the characteristics of housing and neighborhoods are fixed, consumer location and rent patterns provide sufficient information to estimate utility parameters. These are then used to determine a long-run locational equilibrium of the Alonso type, where housing capital is mobile and household decision making depends only on the tradeoff between land and travel.

In my earlier paper, a sample of several thousand households in the San Francisco Bay Area served as a basis for estimating utility parameters (see Bay Area Transportation Study Commission). The stratification of households by income, life cycle, size, race, and occupation resulted in small sample sizes, so parameters were eventually estimated for only seven of the more populous strata. A description of strata characteristics is found in the first four lines of Table 1. The influence of income on loca-

[2] u^o can be replaced in (11) by the expression $\beta_y y^{\alpha_y}$, and the parameters β_y and α_y estimated along with the rest.

tional preferences, when demographic characteristics are fixed, is discernable by examining strata 1 through 4. The final three strata exhibit variation in household size and life cycle when income is constant.

In addition to unit lot size, the housing attributes available in the sample included: number of rooms, age, and condition. Besides the travel time and cost to the respondent's primary work place, locational information included the proximity of the area to retail services and the distribution of income in both the unit's neighborhood and encompassing political jurisdiction. These latter variables should implicitly account for various fiscal and social amenities.

It is beyond the scope of this study, and not really its purpose, to present a detailed discussion of the estimated parameters. Most were significant, and readers interested in the methodology should refer to the earlier research. The intent of this paper is to determine what these parameters imply about income and long-run locational patterns.

Turning to Table 1, the first five lines describe the seven household strata along with the R^2 values for the estimated bid rent functions. Using the parameters of this function, which are in effect utility parameters, lines 6 and 7 compute the marginal expenditure, or income value, of one hour saved in daily travel time.[3] Since "other" expenditure x is in annual dollars, these figures represent the value of an hours commuting each day for a full work year. In keeping with the convexity of indifference surfaces, greater commuting time always has increasing marginal value.[4]

Examining the differences across income groups, a 400 percent increase in earnings between strata 1 and 4 raises the value of commuting time by anywhere from 1,000 to 3,000 percent. Clearly, the distaste for commuting is quite income elastic. What is perhaps more important, however, is that despite such an income elasticity, the value of commuting time is quite small in absolute terms. For the lowest income group it averages around 15¢ per hour[5] or roughly 5 percent of the wage rate. This increases until at the upper income bracket the value is $2.15 per hour, or 12 percent of the wage rate.

While values for commuters' time in this range may seem low, the figures here are actually fairly comparable with estimates developed in the transportation literature. Anne Friedlaender's 1965 study of the interstate highway system valued time at 2.5¢ per mile or around $1 per hour. Several years later Thomas Domencich and Gerald Kraft estimated auto time and cost demand elasticities, which in turn implied a "cost equivalent" value of time that was only 80¢ per hour.[6] More recent studies by Charles River Associates in Pittsburgh and Los Angeles found similarly low cost equivalent figures around $1.20 per hour. In fact, the only study to produce different figures from these is by Daniel McFadden. McFadden's micromodel estimated the value of commuting time as a constant fraction of the wage rate—around 20 to 30 percent. While these numbers are almost twice those developed here, it must be remembered that his model constrained the value to be a constant fraction of income, while the model here takes this as problematic. It is also important to note that during simulation the demand elasticities which emerged from McFadden's micromodel implied a smaller value for commuting time, again only around $1.00 per hour.

Holding income constant, Table 1 suggests that social characteristics also play a role in determining one's preference for commuting. While family size appears to

[3]From (9) this marginal rate of substitution is

$$-\frac{\partial u}{\partial T} \Big/ \frac{\partial u}{\partial x} = -\frac{\alpha_T \beta_T x^{\alpha_o + 1}}{\alpha_o T^{\alpha_T + 1}}$$

where α_T, β_T, and α_o are estimated from (11).

[4]This is consistent with theory if the wage rate is fixed or at least independent of commuting.

[5]Obtained by dividing the marginal rates of substitution in Table 1 by an assumed 300 commuting days per year.

[6]Given time and cost elasticities, E_T, E_c, with respect to auto demand (u), the cost equivalent value of time is:

$$\frac{dc}{dT} = \frac{\partial u}{\partial T} \Big/ \frac{\partial u}{\partial c} = E_T c / E_c T$$

have no effect (stratum 7 versus 2), both older and younger households do have a substantially greater value for commuting time (strata 5 and 7 versus 1).

The second determinant of bid price slopes is the income elasticity of land demand, and moving to lines 8 and 9 of Table 1, marginal rates of substitution between land and "other" goods have been computed for each stratum. For a fixed level of land consumption, these vary by 100 percent between strata 1 and 4. Not surprisingly, this results in a comparable increase in land consumption when location is fixed 15 minutes from work and land is priced at $20,000 per acre (line 10). The computed income elasticity of land consumption, then, is around .25. Unfortunately, there are no published estimates of land demand with which to compare these results. The only related estimate would be the aggregate income elasticity of total housing consumption, most recently established at around .7 (see Garrett Vaughn; R. K. Wilkinson). The difference between these figures may be attributable to the fact that the land estimates here exclude any changes in location with income. If the rich live further out, then the spatial "price effect" will widen the differences in land consumption. Aggregate studies ignore the spatial dimension and hence capture this under an "income effect."

Table 1 also reveals important differences in land demands by social characteristics. Younger households (income constant) have a noticeably lower land demand, and older residents have a slightly higher preference. Finally, greater family size significantly increases the desire for a larger lot.

The next to last line in Table 1 is the most important for the discussion here. Computed from equation (4), it represents the distance derivative of each stratum's bid price gradient—evaluated 4 miles from the city center and assuming each household group is at a level of utility where their bid prices intersect at $20,000. Stratum 6, composed of young people entering or not yet in family formation, clearly dominates all other households with the steepest bid price curve. This results from their high distaste for commuting and an extremely low demand for land. Furthermore, it implies that they should locate most centrally in a simulation of competitive bidding. Poor and moderate income households (strata 1 and 2) have the next most steeply sloped bids and presumably will locate just outside of the urban center. Wealthy households (strata 3 and 4), older families whose children have left (stratum 6), and large families (stratum 7) have slightly flatter bid gradients and should locate more peripherally. The differences in bid slopes among these latter groups, however, all seem quite small in comparison with that of stratum 5.

Comparing the first four strata in more detail, rising income is seen to decrease the slope of household bids very slightly, although the relationship is not completely monotonic. The wealthiest households (stratum 4) have bids that are somewhat steeper than those of stratum 3, but not as steep as strata 1 and 2. In terms of percentages, these differences are all quite small, however, since a 400 percent increase in income reduces bid slopes by only 10 percent.

At first glance, these results seem somewhat difficult to explain. The demand for land is certainly inelastic (.25) while the value of commuting time is quite income elastic (3.0–5.0). From equation (5) one should therefore expect the time effect to dominate and lead to sharply steeper bids with greater income. The reason why this is not the case is that the slope of household bids depends on how the income elasticity of land compares with the income elasticity of *total* travel cost—not just time costs. When the direct money costs of travel are added to the time costs, the result is that total transport costs have an income elasticity of around .25—very close to that for land.

To see this, it must be recalled that in the original bid rent estimating equations (11), the money costs of travel were assumed to be around 6¢ per mile (1965 dollars). With about 300 round trips per year, a move one mile further from the center results in additional annual transportation outlays of $36.

If highway speeds average 25–30 mph, then this additional round trip mile of commuting uses up approximately .07 hours. According to line 6 of Table 1, this is annually worth anywhere from $1.00 to $39.00 (strata 1 and 4 at T = .5). The total annual cost of a mile of travel to stratum 4 is thus $75 or about twice that of stratum 1 ($37). This almost exactly matches the increase in land consumption between these groups.

It is important to note that quite similar conclusions are reached when McFadden's cost equivalent values for time are substituted for those of this study. The final line in Table 1 computes the slope of each stratum's bid price profile under the assumption that the value of commuting time is 25 percent of the appropriate wage rate. Since these values of commuting time are higher than those previously used, it is not surprising to find that all bids are more steeply sloped. The interesting result, however, is that across income groups the differences continue to be surprisingly small. All of this suggests that income in fact may not be a strong determinant of long-run location patterns.

Any such conclusions at this point would have to be considered mostly speculative. The relative bid price slopes of different strata, when evaluated at a common arbitrary point (4 miles, $20,000), are only suggestive of the solution that will prevail in a general equilibrium when bid price curves intersect at different points. That solution itself must be obtained to determine the results which hold when land demands exactly balance the supplies obtained through competitive bidding.

IV

To simulate a long-run equilibrium in the urban land market requires additional system parameters besides consumer utility functions. For the first of these, the circumference of the city was assumed two-thirds circular to represent the loss of land to water in the San Francisco region. All employment was centrally located and land around this point was divided into small rings, each .2 miles in width. Airline distance from each ring to the center was then converted to travel time, with an assumed velocity relationship. Central speeds averaged 10 mph, while at the fringe they reached 50 mph. The opportunity or rural price of land was set at $5000/acre, and in keeping with the estimated bid rent functions, the cost of travel was 6¢ per mile. A linear function was used for the annualized construction cost of housing with a fixed component of $1000, and a variable cost per room of $200. Finally, the total population of the simulated city was set at slightly over one million households, and was partitioned into each of the seven strata in proportion to that group's current representation in the region.

In order to determine the competitive bids of different households, all potential housing in the city was assumed to be "new" and of "sound" quality. Thus, consumer preferences for these attributes, estimated in the original bid functions, would exercise no influence over long-run locational choice. In a similar manner, to stay within the Alonso-Muth framework, social and fiscal amenities were also assumed to be uniform throughout the region. Strata preferences for externalities would therefore also play no role in locational decision making. Given initial utility levels, the marginal conditions (1)–(3) were solved to determine each stratum's bid price, land, and housing consumption in every ring. All land within each ring was allotted to that group which offered the highest net bid price. These land areas were converted into a population "holding capacity" by dividing with the "winning" stratum's land consumption in that ring. Aggregating over those rings commanded by each group yielded an aggregate holding capacity for each stratum, based on the structure of bidding that followed from the initial utility levels.

The general equilibrium conditions (7) and (8) require that these holding capacities (labeled D_i) exactly equal the specified number of households in each stratum

TABLE 2—EQUILIBRIUM LOCATION PATTERNS IN A LONG-RUN COMPETITIVE LAND MARKET

Households Ranked by Location Order	Simulation			
	1	2	3	4
Household/Inner Ring	5/0	5/0	5/0	5/0
Price	66,300	68,000	69,000	87,500
Land	.018	.018	.021	010
Household/Inner Ring	2/2.4	2/2.8	7/3.0	2/2.0
Price	28,900	28,600	26,400	32,000
Land	.114	.116	.152	.061
Household/Inner Ring	1/6.0	4/5.8	1/6.8	4/3.8
Price	18,200	20,500	17,600	24,700
Land	.178	.33	.215	.12
Household/Inner Ring	7/8.6	7/9 0	3/9.8	1/5.4
Price	13,500	13,200	12,850	18,200
Land	.227	.28	.372	.143
Household/Inner Ring	3/11.2	1/10.6	2/12.6	6/7 0
Price	10,100	9,500	10,100	14,000
Land	.431	.31	.45	.192
Household/Inner Ring	6/13.8	3/13.2	6/14 0	7/8 4
Price	7,700	8,000	8,900	11,900
Land	.498	.52	.56	32
Household/Inner Ring	4/15.4	6/16.0	4/16.2	3/12.4
Price	6,480	5,700	7,280	7,100
Land	.980	65	1.60	47
Urban Border	17.8	17.2	21.4	15.6

Note: Household (stratum)/Inner Ring (miles); Price (dollars); Land (acres)

i (N_i). If group i's holding capacity is less than N_i, its utility level (u_i) must be lowered in order to raise its bid price gradient and "capture" more land. The reverse holds as well and so to obtain equilibrium utility levels, the iterative process (12) was used:

$$(12)\quad u_i^t = u_i^{t-1} + \alpha_i[D_i^t - N_i] \qquad i = 1,7$$

At iteration $t - 1$, utility level u_i^{t-1} generates the holding capacity D_i^{t-1} which then is used to adjust utility to u_i^t. Experiments revealed that a fairly wide range of values for the adjustment coefficients (α_i) would lead to convergence over all seven strata. The results of the initial equilibrium solution are shown in Table 2, as simulation 1.

In Table 2 the inner ring of each household represents the radius or distance at which that stratum begins its locational "band." Thus, stratum 5 occupies all land from the city center to a distance of 2.4 miles, while stratum 2 occupies land from that point outward to a distance of 6.0 miles. At the beginning of each stratum's band, the bid price and land consumption for that stratum are also reported. The final line gives the urban border, or the point where the bid of the outermost locating household equals the opportunity price of urban land.

Examining the first simulation, stratum 5 locates centrally as would be expected from the analysis of bid price slopes in Table 1. Its sharply steeper bid function gives it a clear domination over the closest locations. Among the first four strata, rising income is seen at least partially to result in more distant locations. The one exception is stratum 2 which locates closer than stratum 1. This conclusion is somewhat consistent with Table 1. There, the anticipated location order was (closest to farthest) 2, 1, 4, 3. The equilibrium order is: 2, 1, 3, 4. Finally, strata 6 and 7 locate at relatively intermediate distances whereas Table 1 would suggest very peripheral sites. Of course, the

TABLE 3—STRATA BID PRICE GRADIENTS AT INITIAL EQUILIBRIUM

	Strata						
Location	5	2	1	7	3	6	4
1.0 (mile)	48,000[a]	34,200	31,900	27,800	28,300	26,400	30,000
4.0	15,500	23,100	22,800	21,200	20,500	19,800	20,200
7.0	5,300	16,000	16,200	15,900	15,300	14,800	14,700
10.0	620	11,000	11,400	11,500	11,300	11,000	11,010
12.0	[b]	8,600	8,900	9,000	9,150	9,100	9,050
14.2	-	6,300	6,600	6,300	7,100	7,150	7,100
16.4	-	4,900	5,100	4,400	5,600	5,700	5,800

[a]Dollars per acre.
[b]Bid less than zero

reason that Table 1 imperfectly predicts the location pattern is that it compares bid slopes *ex ante*, at a single and somewhat arbitrary intersection point. Such a comparison is in fact valid only *ex post*, at the true equilibrium boundaries. At first glance, then, the results of the initial simulation suggest that greater income may lead to more distant locations in a long-run competitive land market. Unfortunately these results do not hold up with any statistical rigor. That is to say, they are largely the product of chance.

Refering back to the discussion in Section III, Table 1 showed that the differences in bid price slopes among the first four strata were quite small –on the order of only 10 percent. This, it turned out, was due to the fact that the income elasticity of total travel costs was approximately the same as that for land. This suggests that the bid price gradients which produced the simulation in Table 2 may also be quite similar. To determine this, Table 3 reproduces the equilibrium gradients of each strata.

If stratum 5 is excluded for the moment, the variation in bid prices ranges from 20 percent at the city center and urban border to only 4 percent at a distance of 10 miles. This seems quite small, especially when these strata are compared with the bid structure of younger households in stratum 5. What is equally important is that the highest or winning price at a particular location is at most a percentage point greater than the bid of the next most competitive household. Thus, with the exception of stratum 5's central location, the entire equilibrium pattern seems to result from very small differences in bid price gradients. Clearly, it is necessary to determine how robust these differences are with respect to the various parameters that determine them.

The first group of parameters tested included those that do not influence *relative* consumer preferences for location or land consumption. The housing cost function, total urban population, and the opportunity price of urban land are all such parameters, and reasonable perturbations of their initial magnitudes did not affect the locational order of households. To be sure, increasing the fixed cost of housing from $10,000 to $15,000 and raising the opportunity price of land from $5,000 to $10,000 each resulted in more compact, higher density equilibrium cities. The location ranking of households, however, remained the same. In a similar manner, reducing the urban population by 50 percent created a smaller, lower density city, but did not alter relative location patterns.

The second group of parameters included those which could affect the relative bids of consumers: the money cost of travel, the speed or velocity function assumed, and of course the utility parameters of each stratum. Increasing the money cost of travel from 6 to 10¢ per mile increased central rents, but did not alter the pattern of location. The discussion in the previous section

suggests that because a higher money cost of travel decreases the income elasticity of total travel costs, it should enhance the tendency for wealthier households to locate further from the center. Since the initial simulation resulted in this arrangement to begin with, increasing the cost of travel further had no additional impact on relative household location. On the other hand, decreasing the money cost of travel from 6 to 3¢ caused quite a significant change in the equilibrium solution. By allowing the income elastic disutility of travel to play a more dominant role, a lower money cost of travel almost reversed the location pattern of the original simulation. The resultant equilibrium is found in Table 2, as simulation 2. Reducing the speed of travel had similar results. In the original simulation, the speed of travel was assumed to start at 10 mph in the city center, reaching 45 mph at distances greater than eight miles. As an alternative, a uniform speed of 25 mph was tried. The slower speed in the suburbs increased the steepness of wealthy household bids, while faster city speeds had the effect of flattening the relevant portions of low income bid gradients. The result was an equilibrium almost identical to the second simulation in Table 2, one where the location order is close to being reversed. Thus, it seems safe to say that the spatial income gradient produced by this study is indeed sensitive to the assumed characteristics of urban travel.

Perhaps the most important test of robustness is whether the simulation is sensitive to small perturbations in consumer utility parameters. Since these are stochastic variables, fluctuation within one standard deviation of their value could occur by chance. If changes of this magnitude produce alternative location patterns, then these patterns may be largely random phenomena. To test the statistical robustness of the locational equilibrium, two simulations were developed in which each utility parameter was perturbed slightly. Since the variance-covariance matrix of the estimated parameters had significant nondiagonal elements, the pattern of perturbation was not purely random. The estimate of each α in the utility function (9) was correlated (negatively) with its corresponding β parameter. Thus in each test, if an α was perturbed upward, the β was moved downward, and vice versa. As between commodities and households, the perturbations were random—and always less than 1/2 of each parameter's standard deviation. Variation of this small a magnitude is highly probable, due to chance alone.

The resulting equilibria are found in Table 2 as the third and fourth simulations. The locating arrangement of each is quite different from the other, and both in turn are different from the initial solution. With the exception of stratum 5's centrality, the locational order of the first equilibrium is simply not sustained under reasonable parameter perturbation. Very little confidence can be placed in the original conclusion that greater income will even partially lead to more distant residential locations.

V

It is perhaps unusual to think of insignificant results as having significant consequences, yet in many respects that is precisely the implication of this study. It is the view of some urban theorists that households of greater income select more distant suburban locations as a natural consequence of long-run spatial competition. For this to be the case, the income elasticity of land consumption must exceed the income elasticity of the cost of travel—including the value of commuting time. Based on cross-section data, the results of this study strongly suggest that these two elasticities are very similar, in fact so much so that the spatial bidding for land of different income groups looks almost identical. The locational equilibrium that results from these differences in bidding is not at all robust. Its sensitivity to changes in utility parameters, small enough to occur by chance, indicates that any location patterns

by income that results is in some sense statistically unreliable.

The indirect implication of this result is that the long-run spatial theory of Alonso, Mills, and Muth empirically contributes little to the explanation of American location-income patterns. This lends strong credence to the view of other urban economists that the suburbanization of America's middle and upper classes is a response to housing market externalities and the fiscal incentives of municipal fragmentation.

REFERENCES

William Alonso, *Location and Land Use*, Cambridge 1964.

M. J. Beckmann, "On the Distribution of Urban Rent and Density," *J. Econ. Theory*, June 1969, *1*, 60–67.

D. Bradford and H. Kelejian, "An Econometric Model of the Flight to the Suburbs," *J. Polit. Econ.*, May 1973, *81*, 566–90.

A. J. Catanese, "Home and Workplace Separation in Four Urban Regions," *J. Amer. Inst. Plan.*, Sept. 1973, *37*, 331–37.

T. Domencich and G. Kraft, "Estimation of Urban Passenger Travel Behavior," *Highway Res. Rec.*, *238*, 1968, 64–78.

Anne F. Friedlaender, *The Interstate Highway System*, Amsterdam 1965.

Benjamin Harrison, *Urban Economic Development*, Washington 1974.

R. Haugen and A. J. Heins, "A Market Separation Theory of Rent Differentials in Metropolitan Areas," *Quart. J. Econ.*, Nov. 1969, *83*, 660–72.

Edgar Hoover and Raymond Vernon, *Anatomy of Metropolis*, Cambridge 1959.

Gregory Ingram et al., *The Detroit Prototype of the NBER Urban Simulation Model*, New York 1972.

D. McFadden, "The Measurement of Urban Travel Demand," *J. Publ. Econ.*, Nov. 1974, *3*, 303–28.

Edwin S. Mills, *Studies in the Structure of the Urban Economy*, Washington 1972.

Richard Muth, *Cities and Housing*, Chicago 1969.

Jerome Rothenberg, *Economic Evaluation of Urban Renewal*, Washington 1967.

Mahlon Straszheim, *An Econometric Analysis of the Urban Housing Market*, New York 1975.

C. Tiebout, "A Pure Theory of Local Public Expenditure," *J. Polit. Econ.*, Oct. 1956, *64*, 140–47.

G. Vaughn, "Sources of Downward Bias in Estimating the Demand Income Elasticity of Urban Housing," *J. Urban Econ.*, Jan. 1976, *3*, 45–56.

Samuel B. Warner, *Streetcar Suburbs*, Cambridge 1962.

W. C. Wheaton, "A Bid Rent Approach to Urban Housing Demand," *J. Urban Econ.*, April 1977, *2*, 15–32.

R. K. Wilkinson, "The Income Elasticity of Demand for Housing," *Oxford Econ. Pap.*, Nov. 1973, *25*, 361–77.

Bay Area Transportation Study Commission, *Home Interview Survey*, Berkeley 1967.

Charles River Associates, *Economic Analysis of Policies for Controlling Automotive Air Pollution in L.A. Region*, Cambridge 1975.

On Donor Sovereignty and United Charities

By FRANKLIN M. FISHER*

There are eight degrees in alms-giving, one lower than the other. ... [The second degree] is giving alms in such a way that the giver and recipient are unknown to each other. ... [One way to accomplish this is by] the donation of money to the charity fund of the Community, to which no contribution should be made unless there is confidence that the administration is honest, prudent, and efficient.

Below this degree is the instance where the donor is aware to whom he is giving the alms, but the recipient is unaware from whom he received them. The great Sages, for example, used to go about secretly throwing money through the doors of the poor. This is quite a proper course to adopt and a great virtue where the administrators of a charity fund are not acting fairly.

Maimonides

I. The Problem

It is common practice for individual charitable organizations to merge their fund raising activities. The United Fund or Community Chest, the United Jewish Appeal, and Federation of Jewish Philanthropies are well-known examples. There are obvious reasons for such mergers. By having a common fund drive, the combined organizations save considerable expenditure of resources which would otherwise be largely duplicative; moreover, prospective donors are saved the annoyance of having more than one solicitor call. For both reasons, the net receipts of the combined charities may well go up.

On the other hand, such charitable combinations impose a hidden cost on their donors. Whereas before the merger a donor could control the separate amounts which he gave to each charity, after the merger he can generally only control the total amount of his gift. The allocation of that total will generally be decided by the merged organizations. If the donor cares about that allocation, he may be less well off than before.

This phenomenon is plainest where the managements of the combined charities set the allocation explicitly; it is likely to be present, however, even where there is some effort made to accommodate donor preferences. Some merged fund drives, for example, allow each donor to specify how his gift is to be allocated. Clearly, if every donor did so, there would be no utility lost from the inability to control such allocations. In practice, however, large numbers of donors do not avail themselves of this opportunity. The result of this is that a large sum of otherwise unallocated money is distributed by bargaining among the managements of the component charities. One very important element of that bargaining inevitably becomes the financial needs of each organization after taking into account the earmarked funds. Hence, if one charity gets a high proportion of the earmarked funds, it is likely to do relatively less well in the later bargaining for the undifferentiated funds than a charity receiving smaller earmarked donations.[1] If donors perceive this to be the case, then they will also perceive that their earmarking does not affect the ultimate disposition of their gifts. (Indeed, this may be one reason for the failure to earmark, although certainly not the only one.)

*Professor of economics, Massachusetts Institute of Technology.

[1]This does not have to be the case, of course. The management of the combined charity, for example, could take earmarking as an expression of donor preferences and divide the undifferentiated funds in the same proportion as the earmarked funds. Such a solution is unlikely to be very stable in practice, however, unless every individual charity gets as much as it would expect to get on its own. In general, bargaining over fair shares is likely to be a complicated business.

Moreover, if the combined fund drive accounts for a large share of the funds raised by the component charities, the same phenomenon can arise even if those charities have supplemental individual fund raising activities. A donor giving to a particular charity may be directly contributing to it but may be weakening that charity's bargaining position in the distribution of the combined charity's receipts. In this case, as in the case of earmarking, if a dollar of direct contribution results in a dollar less of allocation from the combined fund, then donor activities have no effect on ultimate allocation. If the relation is other than one-for-one, then donors can affect the ultimate allocation, but not as efficiently as if the charities were wholly separate. In either case, a utility loss is imposed on the donor.[2]

Do donors care about the allocation of their own funds or only about the existence of the charities involved? It seems plain to me that donors do care about such allocations. Certainly it would be a very strong assumption to suppose that they do not. Unless donors care, it is hard to understand why individuals give more—sometimes considerably more—to some charities than to others. While it is true that charities partake of some aspects of public goods (in the technical sense), it seems clear that donors derive satisfaction not merely from the knowledge that the charities exist but also from the sense of themselves participating in a worthy cause. Unless they think all causes equally worthy, they are likely to care where their money goes.

Note that this means that utility losses can be present even if the management of the combined charities sets the postmerger allocation of funds so that each individual charity's share is the same as it was before the merger. (This may, of course, be a sensible thing to do.) Even though such an allocation is the average, in some appropriately weighted sense, of the allocations which donors would choose, there may be no donor for whom it is the preferred allocation of his own money. A donor who cares about the allocation of his own contribution and not just about the total funds going to each component charity will then be made worse off by being forced to contribute in the average proportion.

Hence, while combined fund drives provide obvious resource savings and may also involve utility gains to donors in the form of decreased annoyance,[3] they are also likely to impose utility losses on donors because of lessened or lost control over fund allocations. Two questions then arise.

First, if donors are unhappy about the way in which their money is to be allocated, are they not likely to express that unhappiness by changing the amounts they give? Since the management of the charities is likely to be sensitive to the total receipts, doesn't this mean that donors, by voting with their dollars so to speak, will influence the allocation in the way they would like it to go? At the least, one would expect this to be true if there is no problem of aggregating over donors with widely different preferences.

Second, if after the merger, receipts net of administrative and fund raising expenses go up, can that not be taken as an indication that the utility costs imposed on donors are more than offset by the resource savings? Certainly, if *gross* receipts go up, one would expect this to be an indication that donors are happier with the merger than they would be without it. Hence one might expect to judge whether the merger was worth having by looking at gross or net receipts.[4]

[2]If donors believe erroneously that they can completely control the allocation of their gifts, is there a *real* utility loss? For my purposes it seems unnecessary to explore this question.

[3]As exemplified by the time and trouble of the "great Sages" described in the opening quotation. Perhaps it is worth remarking, however, that in modern times the merging of charities does not provide the donor with the satisfaction of rising one step up the Ladder of Charity of Maimonides. The feature which distinguishes the second from the third degree is the question of the anonymity of the individual ultimate recipients, and this is generally equally preserved whether or not the charities are merged.

[4]I am well aware that all of this is from the point of view of the donors only. Clearly the merger will be worth having from the point of view of the ultimate recipients if net receipts go up and each component

The present paper shows that both of these suppositions are erroneous as general propositions. I provide a counterexample with a single donor in which the total amount given to the merged charities goes up as the postmerger allocation moves *away* from that which the donor would choose in the absence of the merger. Indeed, for a particular special case, total charitable donations are actually *minimized* at the preferred allocation (at least locally). Hence management paying attention to total receipts will generally be led away from the allocation preferred by the donor. It follows immediately that one cannot conclude the merger was worth having because gross or net receipts go up.[5] Generalization to many donors is immediate.

While the example used is, of course, special, it is in no sense pathological.[6] Indeed, as the later discussion suggests, the results will certainly hold for a wide class of utility functions. Hence, while there may be occasions on which gross (or net) receipts provide an appropriate guide to donor preferences and to the desirability of mergers from the donors' point of view, they do not do so in general.

I now proceed to the mathematics, postponing heuristic discussion until the examples have been given.

II. The Counterexample

Assume a single donor who allocates his income y among donations to two charities, denoted x_1 and x_2, and expenditure on a single ordinary commodity, denoted x_3. We choose the units of x_3 so as to make its price unity. The donor then (with no merger) faces the budget constraint

$$x_1 + x_2 + x_3 = y \tag{1}$$

The donor maximizes a strictly quasi-concave utility function

$$U(x) = V(x_1) + W(x_2) + Q(x_3) \tag{2}$$

where

$$\begin{aligned} V(x_1) &= \beta_1 \log(x_1 - \gamma_1) \\ W(x_2) &= \beta_2 \log(x_2 - \gamma_2) \end{aligned} \tag{3}$$

with the $\beta_i > 0$.[7] There is no need to restrict $Q(x_3)$ further than required for strict quasi concavity. We assume until further notice that

$$\beta_1 \gamma_2 \neq \gamma_2 \beta_1 \tag{4}$$

Counterexamples are permitted to be special, of course; it may be felt, however, that this one is objectionable in a particular way. I have made the donor's utility function depend solely on his own consumption and his own contributions—that is, on his own feeling of contributing to worthy causes. Donors can obviously also be interested in the existence and level of charities independent of their own contributions (the public good aspect of charities referred to above), and the donor whose behavior is being modeled apparently is not interested in such considerations.

Such a defect in the example is apparent

---

charity gets at least as much with the merger as it would without it. It is not at all clear how one should weigh the interests of the recipients against the interests of the donors. Presumably the donors consider the interests of the recipients in making their donations (formally, the utilities or consumptions of the recipients enter the utility functions of the donors). Is one justified in taking further account of recipient utilities than this and imposing on donors some outside sense of what their charitable obligations should be? This is not a simple question and I do not consider it further in this paper.

[5]Of course donors will probably be pleased if all charities get more than before. The issue is whether they will be sufficiently pleased to offset their annoyance at having their allocations restricted and (possibly) paying more than they would like. This depends on the relative importance in the donor's utility function of the donor's own gifts and the total moneys received by the charities.

[6]As would be, for instance, a case in which the donor insisted that the net amount received from his personal donation by a particular individual charity be at least some minimum. In such an example, the donor would obviously regard the imposition of an outside allocation following the merger just as he would an additional administrative expense and feel compelled to give more to achieve the same net result. While instructive, such an example seems too extreme to be persuasive.

[7]It is necessary to assume that $x_1 > \gamma_1$, $x_2 > \gamma_2$ is feasible, which will certainly be true if the $\gamma_i < 0$, but can hold even if the $\gamma_i > 0$ provided income is large enough.

rather than real, however. Let Z_1 and Z_2, respectively, denote the level of donations by all *other* individuals to the two charities. We could replace (2) by taking the donor's utility function to be

$$U^*(x,Z) = F(U(x), Z_1, Z_2) \tag{5}$$

where $U(x)$ is given by (2). Since the Z_i are outside the control of the particular donor, however, we may as well take them as parameters and, given the weak separability of (5), treat the donor as though his utility function were simply (2). While such separability is also special, continuity will show that our results continue to hold when such separability is absent but departures from it sufficiently small. Since the purpose of a counterexample is to place the burden of proof on those believing the propositions being negated, this is sufficient generality for our purposes.

Denoting differentiation by subscripts, the first-order conditions for the premerger optimum are:

$$V_1 = W_2 = Q_3 = -\lambda; \quad x_1 + x_2 + x_3 = y \tag{6}$$

where λ is a Lagrange multiplier.

Now suppose that the two charities merge and allocate their funds so that the ratio of the first charity's funds to those of the second are given by k, a fixed positive constant announced to all donors. Hence $x_1 = kx_2$, and the donor now faces the problem of choosing x_2 and x_3 to maximize

$$U(kx_2, x_2, x_3) = V(kx_2) + W(x_2) + Q(x_3) \tag{7}$$

subject to

$$(k+1)\,x_2 + x_3 = y \tag{8}$$

(Note that the merger changes the relative prices of x_1 and x_2, although that is not the only thing it does.)

Define $c = x_1 + x_2$, the donor's total contributions to charity. If the conjectures we are examining are correct, then such total contributions will be greatest if the merged charity sets k at the ratio which the donor would prefer, the ratio he would himself choose in the absence of the merger. I shall show that this is not the case in the present example by showing that $\partial c/\partial k \neq 0$ at the premerger optimum. It follows that the donor's contributions will *rise* rather than fall as the merged charity's managers move away from his preferred allocation in a particular direction.

The first-order conditions for the postmerger optimum are

$$kV_1 + W_2 + (k+1)\lambda = 0; \quad Q_3 + \lambda = 0; \quad (k+1)x_2 + x_3 = y \tag{9}$$

which involve the intuitive condition that the marginal utility of a dollar consumed equals a weighted average of the marginal utilities of dollars donated to each charity. Differentiating totally with respect to k:

$$\begin{bmatrix} k^2V_{11} + W_{22} & 0 & k+1 \\ 0 & Q_{33} & 1 \\ k+1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \partial x_2/\partial k \\ \partial x_3/\partial k \\ \partial \lambda/\partial k \end{bmatrix} = - \begin{bmatrix} V_1 + kx_2V_{11} + \lambda \\ 0 \\ x_2 \end{bmatrix} \tag{10}$$

Observe that, since income is fixed, $\partial c/\partial k = -\partial x_3/\partial k$. Let D be the determinant of the matrix on the left of (10) and observe that $D > 0$ by the second-order conditions. Inverting that matrix by the adjoint method, we obtain:

$$\begin{aligned} \partial c/\partial k &= -\partial x_3/\partial k = \\ &(1/D)\{(k+1)(V_1 + kx_2V_{11} + \lambda) - x_2(k^2V_{11} + W_{22})\} \\ &= (1/D)\{(k+1)(V_1 - Q_3) + x_1V_{11} - x_2W_{22}\} \end{aligned} \tag{11}$$

using (9) and the fact that $x_1 = kx_2$.

Now consider setting k at the allocation the donor would himself choose, so that the first-order conditions for the premerger optimum (6) are satisfied. At such a point, $V_1 = Q_3$. Further, because of the particular character of V and W given in (3), the con-

dition that $V_1 = W_2$ becomes

$$(12) \qquad \frac{\beta_1}{x_1 - \gamma_1} = \frac{\beta_2}{x_2 - \gamma_2}$$

Hence, at such a point (evaluating V_{11} and W_{22}), (11) becomes

$$(13)$$
$$\partial c/\partial k = (1/D) \left\{\frac{\beta_2 x_2}{(x_2 - \gamma_2)^2} - \frac{\beta_1 x_1}{(x_1 - \gamma_1)^2}\right\}$$
$$= (1/D) \left(\frac{\beta_1}{x_1 - \gamma_1}\right) \cdot \left(\frac{x_2}{x_2 - \gamma_2} - \frac{x_1}{x_1 - \gamma_1}\right)$$

However, from (12)

$$(14) \quad \frac{x_2}{x_2 - \gamma_2} = \frac{\beta_2 x_1 + (\beta_1\gamma_2 - \beta_2\gamma_1)}{\beta_2(x_1 - \gamma_1)}$$

Hence the sign of $\partial c/\partial k$ at the premerger optimum is the same as that of $(\beta_1\gamma_2 - \beta_2\gamma_1)$ and (4) states that the latter magnitude is not zero, yielding the desired result.

I now consider the special case in which the inequality in (4) does not hold and show that there exist subcases in which total donations are actually at a local minimum at the premerger optimum.[8] In order to do this most easily, observe the following properties which hold at the premerger optimum if (4) is violated and the γ_i are not zero.

First, since $\partial c/\partial k = 0$ in such a case, it follows that

$$(15) \quad \partial x_1/\partial k = -\partial x_2/\partial k = x_2/(k+1); \quad \partial x_3/\partial k = 0$$

Next, from (14) in this case, it must be true that

$$(16) \qquad k = x_1/x_2 = \gamma_1/\gamma_2$$

Denote by N the term in brackets on the far right-hand side of (11) and note that $N = 0$ at the premerger optimum in this case as does $(V_1 - Q_3)$.

At the premerger optimum in this case, therefore,

$$(17) \quad \partial^2 c/\partial k^2 = (1/D)\partial N/\partial k = (1/D)\{(k+1)V_{11} + V_{11} + x_1 V_{111} + W_{22} + x_2 W_{222}\}\left(\frac{x_2}{k+1}\right)$$

where use has been made of (15). From the definition of V, however,

$$(18)$$
$$V_{11} + x_1 V_{111} = \frac{\beta_1}{(x_1 - \gamma_1)^2}\left(\frac{2x_1}{x_1 - \gamma_1} - 1\right) = \frac{\beta_1}{(x_1 - \gamma_1)^2}\left(\frac{x_1 + \gamma_1}{x_1 - \gamma_1}\right)$$

and similarly for W. Using (12), (16), and (18), we obtain from (17):

$$(19) \quad \partial^2 c/\partial k^2 = \left(\frac{1}{D}\right)\left(\frac{x_2}{k+1}\right)\left(\frac{\beta_1}{(x_1 - \gamma_1)^2}\right) \cdot (k+1)\left(\frac{x_1 + \gamma_1}{x_1 - \gamma_1} - 1\right) = \gamma_1\left(\frac{2x_2\beta_1}{D(x_1 - \gamma_1)^3}\right)$$

which has the sign of γ_1.

Thus, while in the very special case we are examining, total donations are maximized (at least locally) at the preferred allocation if the γ_i are negative, they are actually at a relative minimum in the equally plausible case in which the γ_i are positive.[9] As for the case in which both γ_i are zero (thus still violating (4)), as one might expect, this turns out to be very much a watershed. Indeed, in this case, total donations turn out to be wholly independent of the allocation set by the merged charity. To see this, observe that in this case the postmerger first-order conditions (9) imply

$$(20) \qquad c = (k+1)x_2 = \frac{\beta_1 + \beta_2}{Q_3}$$

so that x_3 must satisfy

[8] I suspect that in some (or all) of these the minimum is global, but this is harder to show.

[9] Curiously, this turns out to be the case even though, in view of (12) and (16), the premerger allocation is independent of income so that the donor seems especially attached to it in some sense.

$$x_3 + \frac{\beta_1 + \beta_2}{Q_3} = y \tag{21}$$

an equation which is independent of k. In view of the budget constraint (1), this means that c is also independent of k.

III. Heuristics

What can be said by way of intuitive discussion of these perhaps surprising results?[10] One way of looking at the matter is to think of the merged charities as a monopolist engaging in a tie-in sale. Without the merger, the monopolist has no control over the prices of the goods he sells, but after the merger he does have some aspects of control over their *relative* prices (see (8)). There is no reason to suppose that it will be optimal in the postmerger situation for the monopolist to set the relative prices as though he had no such control.

While there is a good deal in this, such an explanation ignores the fact that the "monopolist" does not simply set relative prices but instead makes a much more complicated offer. Moreover, this explanation tells us little about the behavior of the donor himself and thus begs the question somewhat.

Turning to donor behavior, it is easy to see (as remarked in an earlier footnote) that certain extreme kinds of utility functions will generate the result obtained. But the utility functions used above are not particularly extreme. What then is going on?

One way of seeing that the result is likely to be true is to generalize the extreme case just referred to as follows. Suppose the donor cares only for Charity A and is indifferent as to Charity B. Before the merger, we can think of him as purchasing "Charity A certificates" at a dollar apiece, where each certificate represents the receipt of one dollar by Charity A. After the merger, the merged charities allocate some fraction of the jointly collected funds, say 10 percent, to Charity B. From the point of view of this particular donor, this amounts to roughly a 10 percent increase in the price of Charity A certificates. We should naturally expect his demand for those certificates to fall. There is no reason, however, why it must fall by as much as 10 percent. If his demand is inelastic, the total dollar donation he makes will rise even though he would prefer an allocation with 100 percent rather than 90 percent going to Charity A.

The important point in considering this still special example is to realize that the thing bought (dollars delivered) and the amount paid (dollars donated) are not the same. With this in mind, one can now go a bit deeper and consider the underlying utility maximizing behavior involved in more general cases. Thus, consider the following argument (which, incidentally, suggests that the result is not at all restricted to the particular utility functions used in the counterexamples mathematically explored above). Consider a donor at the postmerger optimum whose preferred allocation of funds is different from that set by the charity. Clearly, given his consumption and the *total* amount donated to charity, his charitable funds are not being allocated efficiently. Now, with consumption fixed, remove the allocation constraint and permit him to allocate the same total charitable funds in an efficient way. He will do this by reallocating funds from the charity with the lower marginal utility to the charity with the higher marginal utility (see (9)), stopping when the two marginal utilities are equal. At this point, however, the marginal rate of substitution between consumption and charitable donation will have been altered. While it is possible that the marginal utility of consumption is now lower than that of donation (recall that the prices are all unity), there is no reason why this must be so. If it is not, he will wish to transfer funds from donation to consumption, thus giving less to charity at the premerger optimum than at the postmerger one.

Another way of looking at the matter is

[10] For treatment of a somewhat analogous problem, see Edmond Phelps and Robert Pollak.

as follows. With consumption fixed at its postmerger value, consider the minimum expenditure on charity required to reach the postmerger utility level if allocations are not restricted. Since the allocation restriction prevents the donor from achieving his postmerger utility level efficiently, the removal of that restriction will enable him to do just as well with some money left over. Now, it may be optimal for him to spend all the money so saved on charity and even optimal for him to do so and then transfer funds from consumption to charity, but it is not inevitable that he should do so. Looked at in this way, there seems no reason to suppose that he will not wish to spend some of the money saved on consumption, in which case he will give less to charity at the premerger than at the postmerger optimum.

As these arguments suggest, removing the allocation restriction allows the donor to substitute donations to one charity for donations to the other so as to efficiently allocate his charitable funds. In doing so, however, there are effects on the marginal rate of substitution between either charity and ordinary consumption that can go more than one way.

This is the best that I have been able to do in seeking to explain the results. The astute reader will notice that I have not attempted a heuristic explanation of the results for the special cases in which (4) fails to hold and the premerger allocation turns out to be either a local minimum or a local maximum. This is left as an exercise.

IV. Conclusions

Even before aggregation problems, therefore, it turns out that donor sovereignty over charitable allocations is unlikely to occur. Even if the managers of the merged charity pay strict attention to donors and seek to maximize gross receipts, they will not generally be led to the allocation which donors prefer. Indeed, there exist cases in which any move *away* from the donor-preferred allocation increases gross receipts.

A fortiori, it is not the case that one can conclude that such a merger is desirable from the point of view of donors by seeing whether it increases net or gross receipts. Such receipts can go up rather than down just because donors are forced to give in proportions which they do not freely choose.

Indeed, this may be the case even if the managers of the charity set the postmerger allocation equal to that which obtained before the merger (a natural thing to do, but one which will be harder to justify the farther in the past is the original merger). Receipts may then go up not because donors are pleased at being saved one or more solicitations, but because they are forced to give in the average proportions even though every one of them feels worse off as a result.[11]

The interesting questions of when such mergers are desirable and how the allocations should be set must therefore be examined according to other criteria.

[11]To see that such an example is possible, suppose that donors fall into two classes. Let every donor in the first class have a utility function of the partial Cobb-Douglas type discussed at the end of the preceding section, that is, with both $\gamma_i = 0$. Let the other class all have utility functions of the type described in the more general counterexample. It is easy to see that if the first group has a higher than average preferred value of k and the second group a lower, then total receipts can increase when everyone is forced to the average allocation. This will occur because the donations of the first group will remain unchanged while those of the second group will increase (for appropriate choice of the parameters).

REFERENCE

E. S. Phelps and R. A. Pollak, "On Second-Best National Saving and Game-Equilibrium Growth," *Rev. Econ. Stud.*, Apr. 1968, *35*, 185–200.

Do Schools Make a Difference?

By ANITA A. SUMMERS AND BARBARA L. WOLFE*

Parents, courts, and legislatures have been struggling to define equal educational opportunity (minimum achievement level for all? minimum growth in achievement? differential growths in achievements?). At the same time, economists, sociologists, and educators have been struggling to identify which package of school inputs is required for each type of student to equip him or her for educational growth. Most empirical attempts to identify which inputs matter have concluded that schools barely make a difference. From this conclusion has flowed a prevailing nihilism with respect to schools as an egalitarian force. We conclude, on the basis of a microeconometric examination of Philadelphia School District data, 1) that many school inputs do matter, 2) that disadvantaged students can be helped by particular types of inputs, and 3) that the use of pupil-specific data, and statistical methods appropriate to such data, account for the cheerier results of this study.

Little theory, economic or otherwise, is currently available to describe the determinants of educational achievement. Casual observation, combined with the education literature, suggests that achievement (A) is a function of a student's hard-to-disentangle genetic endowment and socioeconomic status ($GSES$), teacher quality (TQ), non-teacher school quality (SQ), and peer group characteristics (PG). Thus,

$$(1) \qquad \Delta A = F(GSES, TQ, SQ, PG)$$

This equation is, of course, a reduced form. Attitudinal variables may affect achievement, but are themselves determined by the four factors listed above.

Much of the empirical literature that has arisen in the years following the Coleman Report (see James S. Coleman et al.) has focussed on estimating equation (1). Many investigators have viewed it as a production function relationship, describing educational output as a function of inputs. In education, all inputs cannot be selected as in a factory. Such a view, therefore, blurs the distinction between the variables which the educational policymaker can control (TQ, SQ, some PG) and those which he cannot ($GSES$, some PG). Further, the production function, as used in its classical context, relates the maximum attainable level of output for given inputs to the level of inputs—it describes the boundary of the production set. There is little reason to believe that we now know enough to have any confidence at all that schools are attaining such productive efficiency. In any case, it is clear that estimation procedures based upon cost-minimization assumptions are inappropriate. It seems preferable, therefore, to view (1) as a simple input-output relationship.

Past attempts at estimating (1) have represented many inputs by school- or district-wide averages, rather than by the more appropriate pupil-specific data;[1] we use pupil-specific proxies for the variables and find that much can be revealed by disag-

*Research officer and economist, Federal Reserve Bank of Philadelphia; consultant, Federal Reserve Bank of Philadelphia and economist, Institute for Research on Poverty, University of Wisconsin, respectively. We are indebted to Robert Summers and Lawrence H. Summers for their suggestions and interest, and to the School District of Philadelphia for their willingness to allow us access to their detailed data.

[1]School inputs were measured at the district level in the studies by Charles Benson et al., Elchanan Cohn, and Herbert J. Kiesling; they were measured at the school level in the studies by Harvey A. Averch and Kiesling, Coleman, Jesse Burkhead et al., and Martin T. Katzman; the studies by Samuel Bowles, Eric Hanushek, Henry M. Levin, and Stephan Michelson, which drew upon Coleman's *EEO* data, also used school inputs aggregated to the school level. Hanushek (1970), Richard Murnane, and Donald Winkler used some pupil-specific data.

gregation.[2] Past studies have used an achievement level as the output measure;[3] we use a value-added dependent variable—the change in achievement over a period of years. Finally, past studies have either relied on an additive form, or have segmented the sample, and estimated, conditional on *GSES*;[4] we generalize these procedures by using interaction terms to allow the effect of policy variables (*TQ*, *SQ*, *PG*) to vary with *GSES*.

We conclude that empirical investigations have failed to find potent school effects because the aggregative nature of the data used disguised the school's true impact. The use of school and district averages introduced so much noise as proxies for inputs into students that effective inputs were not revealed. The use of achievement level as an output measure without controlling for the beginning level, clouded the examination of school inputs on that which schools can impact. And the lack of availability of pupil-specific data prevented an adequate exploration of the differential impacts of *TQ*, *SQ*, and *PG* on students with different *GSES*—and, in fact, in several instances have completely hidden the input effects.

Section I describes the sample and model estimation procedures. The major findings and caveats are in Section II. The importance of examining school inputs differentially is established in Section III by a comparison of the results with and without interaction analysis. Section IV documents the importance of using disaggregated data by a comparison of the results when pupil-specific data for school inputs are used with the results when school averages are used. Some concluding remarks constitute Section V.

I. Description of Sample and Model Estimation Procedures

We had access to an exceptionally rich data base, a necessary condition to test the hypothesis that disaggregation would reveal more significant inputs. This study is based upon data from 1970-71 pupil files for 627 sixth-grade elementary school students in 103 elementary schools. Schools were selected randomly from the Philadelphia School District, and students were randomly selected from their schools. Data were similarly gathered and analyzed for 553 eighth-grade students in 42 junior high, K-8, and middle schools, and for 716 twelfth-grade senior high school students in 5 senior high schools. Only the elementary school data and results are discussed in the body of this article.[5] Analysis of the eighth-grade sample produced similar results—specific comparisons are noted in several footnotes. The twelfth-grade sample drew upon so few schools that the results were of limited interest.

A three-year personal educational history was compiled for each student. This was then matched with data on school-wide resources of the school the pupil attended, with the pupil's estimated family income,[6]

[2]In another study which analyzed the intradistrict distribution of school resources to disadvantaged students in Philadelphia, examination of detailed individual school data on each of twenty-six to thirty-one school inputs revealed useful policy insights. For example, while the instructional cost per pupil (a commonly used aggregated measure) was higher in schools with higher proportions of black and low income pupils, these schools were in general staffed by teachers with less experience and less education beyond the B.A. and by teachers who came from lower rated colleges and had scored lower on the National Teacher's Examination. Common. (See the authors, 1976.)

[3]All of the studies cited in fn. 1 used achievement as the dependent variable.

[4]Averch and Kiesling, Burkhead et al., Cohn, Katzman, and Levin used the additive form; Benson et al, Coleman (1966), Hanushek (1970, 1972), Kiesling, Michelson, and Winkler used subsamples; Murnane examined the interaction effect of achievement level.

[5]See the authors (1974) for the analysis and results of the eighth- and twelfth-grade samples, and for more detail on the sixth-grade sample.

[6]We have developed a procedure using 1970 Philadelphia Census data for estimating block income from block mean housing values, block mean contract rental values, tract distribution of block contract rental values, and tract distribution of income values. The block income appropriate to each pupil was taken to be his or her family income. This procedure involves 1) forming the cumulative distributions of data for each tract of owner-occupied housing values, contract rental values, and family income; 2) converting these cumulative distributions into relative distributions

and with data on each pupil's individual teachers. It was possible therefore to 1) look at pupils longitudinally, 2) examine a great many variables in a pupil-specific way (the teacher variables are of particular interest), and 3) go beyond simple linear specifications because of the fairly large number of observations.

The dependent variable chosen is the change in a composite achievement score—achieving—over the three-year period, third to sixth grades.[7] Using the final score, which does not visibly control for initial achievement level, as the impact measure of resources is less satisfactory. This use of a value-added measure is consistent with the usual choice in estimating production functions. Further, the change formulation permits the prediction of the effect on pupil learning of educational input changes.[8]

The independent variables were of three types: the genetic and socioeconomic character of the pupil (for example, family income, IQ, race), pupil-specific school inputs (for example, size of class of pupil's grade, qualities of the teachers who taught the pupil), and peer group characteristics (for example, proportion of high achievers, proportion of blacks). The relationship was examined using single equation multiple regression. Dummy variables, piece-wise linear fitting, and other nonlinearities were employed. Interactions of income, race, and/or achievement with school input and peer group variables were explicitly introduced.

Past attempts to estimate education production functions using a single equation regression method have been criticized on the grounds that they have ignored important problems of simultaneity. The argument, in particular, is that achievement gains affect student attitudes, and student attitudes affect achievement gains. Thus variables appearing on the right-hand side and treated as exogenous may in fact be endogenous. In this study, however, very few "psychological" variables are used. Only two of the variables might conceivably be regarded as problems. They measure the student's attendance record, and are included to reflect the student's motivation. However, reverse causality flowing from achievement (as measured by tests) to motivation seems to be tenuous: children are frequently unaware of their achievement test scores; grades and teacher attitudes are probably far more important in attitude reinforcement than Iowa scores; and there is considerable evidence that grades are poorly correlated with standard test scores. When the equations were estimated in the reduced form version, without the attendance variables, the coefficients were not changed in a way which altered any conclusions.[9]

(percentiles); 3) determining for each block the percentile in the tract distribution of mean owner-occupied housing value and the percentile for mean contract rental value; 4) determining the corresponding normal deviate arguments, 5) adjusting these by the regression coefficient for the tract between housing and income data for a cross-classified 20 percent sample; 6) assigning percentiles to the adjusted arguments; 7) finding the income values for these percentiles, 8) adjusting for differences in the income distribution of renters and owners, 9) averaging the two income values for each block. The procedure was carried out for black and nonblack income and housing distributions, and each pupil was assigned an income estimate on the basis of his or her race. See the authors (1975).

[7]The achievement score was measured in terms of grade equivalents—the norm for a sixth-grade student, tested in April of the sixth-grade year, would be 6.8, for example. For the students in the sample, the average sixth-grade score was 5.2. The grade equivalent form was preferable to either of the available alternative forms. the raw scores could not be compared over different grades, since the test differed, and the national percentile scores introduced changes in the nation's student population which were irrelevant.

[8]This formulation, it has been argued, is erroneous because the differences between initial and final score regress to the mean—that is, because tests have random error, there will tend to be a negative correlation between initial score and change in achievement. The concern is, of course, that if initial achievement is omitted from the right-hand side, the estimates of all the coefficients of variables correlated with initial achievement will be biased. However, when the regressions were run with and without the initial score, the variables which were significant remained so, and the changes in the coefficients were in no instance large enough to alter any of the broad conclusions drawn from the study.

[9]See the authors (1974, p. 54).

II. Summary of Findings

In winnowing down the original list of variables to get the equation of "best fit," many regressions have been run. A variety of functional forms—non-linear, interactive—for the variables have been examined. The data have been mined, of course. One starts with so few hypotheses convincingly turned up by theory that classical hypothesis testing is in this application sterile. The data are there to be looked at for what they can reveal.

The standard tests of significance provide guidance of only a very crude sort—hence, the usual asterisks are misssing from the t-statistics in the tables. We threaded our way through the myriad of variable combinations with the magic 5 percent normal curve numbers as our North Star. But, to the final formulations and interpretations of coefficients, more informal standards were applied. All interaction results, for example, were checked out against separate regressions based upon data subsamples. Variables which had coefficients whose significance were very sensitive to the introduction and discarding of other variables were not retained. All the major findings were extensively tested for robustness to alternative specifications.

The coefficients and t-statistics[10] for the equation of best fit for the sixth-grade sample are shown in column (2) of Table 1. The definitions, sources, $\bar{X}$'s, and σ's of the variables are listed in the Appendix.

We are confident that the coefficients describe in a reasonable way the relationship between achieving and *GSES*, *TQ*, *SQ*, and *PG* for this collection of 627 elementary school students in the Philadelphia School District. What we do have to be agnostic about is the absence of the results of replication.[11]

A. *Genetic and Socioeconomic Inputs*

In addition to examining the effects of Income (row 1) and Race (row 2) on learning, interactions between school inputs and income and race of pupil were carefully examined—and a great deal of interaction was revealed. In fact, when these interactions are accounted for, no residual impact of Race and Income on achievement growth remained. The Sex (row 3) of the student affected learning—males did worse than females. Over the three-year period, males grew almost one month less in achievement scores than females.

The first-grade IQ (rows 4, 5, and 6) of the pupil had a strong effect on achievement growth. But an examination of the interaction of IQ with Race for high IQ students revealed that there was an additional positive effect for nonblack students only. That is, nonblack students with high IQ's grew more rapidly than those with lesser IQ's (2.8 months, compared with 1.3 months over the three years for each 10 point increment in the IQ score), but black students experienced similar achievement growths regardless of IQ. (It might be argued that, by controlling for initial ability, one channel through which *GES* might affect achievement is eliminated. When the best fit equations were estimated without IQ, however, the coefficients related to Race and Income changed very little.)[12]

The motivation of students (proxied by Unexcused Absences and Latenesses) had a significant bearing on learning. Students with more Unexcused Absences (rows 7 and 8) and more Latenesses (rows 9 and 10)

[10] For interaction variables, the t-statistics in the table indicate only whether or not there is a significant difference in the impact of each input among different types of students. Tests of the form.

$$\bar{\theta} = \frac{\hat{\alpha} + \hat{\beta} y^*}{\sqrt{\hat{\sigma}_{\hat{\alpha}}^2 + \hat{\sigma}_{\hat{\beta}}^2 y^{*2} + 2\hat{\sigma}_{\hat{\alpha}\hat{\beta}} y^*}}$$

were used to determine the specific values of the interaction terms at which the results were significant. The interesting conclusion was that a considerable amount of variation over the range was revealed.

[11] We are now in the midst of another phase of this study using a much larger sample of data from the same period (several thousand for each level of schooling). While it is regretful that these data are less pupil-specific, it enables the examination of the losses due to disaggregation in considerable detail.

[12] See the authors (1974, p. 51).

TABLE 1—REGRESSION RESULTS FOR SIXTH-GRADE SAMPLE OF PHILADELPHIA SCHOOL DISTRICT STUDENTS, 1968–71[c] (t-values in parentheses)

	Pupil-Specific School Inputs		School Averages of School Inputs	
Variables	With Interaction Terms	Without Interaction Terms	With Interaction Terms[a]	Without Interaction Terms
(1)	(2)	(3)	(4)	(5)
1. Income	.06 (.34)	.17 (1.08)	-.002(-.009)	.21 (1.26)
2. Race	-.23 (-.10)	-3.34 (-2.58)	.74 (.31)	-3.03 (-2.47)
3 Sex	-.90 (-1.39)	-1.23 (-1.88)	-.83 (-1.25)	-.96 (-1.45)
4. IQ	.13 (3.64)	.13 (3.75)	.15 (4.02)	.15 (4.31)
5. IQ 110+	.15 (1.40)	.09 (.98)	.14 (1.32)	.05 (.58)
6. (5) x Race	-.22 (-1.78)	-	-.22 (-1 75)	-
7 Unexcused Absences	.14 (.53)	-.30 (-3.63)	.13 (.47)	-.26 (-3.17)
8. (7) x Income	-0.6 (-1.84)	-	-.05 (-1.54)	-
9. Latenesses	-1.11 (-3.38)	-.20 (-2.00)	-1.20 (-3.45)	-.25 (-2.36)
10 (9) x Income	.11 (2.93)	-	.11 (2.82)	-
11. Rating of Teacher's College, I ≥ 525	13.57 (2.45)	4.03 (2.57)	-2.92(-.55)[b]	-3.35 (-1.68)
12. (11) x Income	-.99 (-1.74)	-	.04 (.08)	-
13 Teacher's Experience	-.47 (-1.80)	.07 (.73)	-.70 (-1.06)	-.05 (-.11)
14 (13) x Third-Grade Score	.02 (2.29)	-	.02 (1.22)	-
15. National Teacher Exam Score	-.02 (-2.62)	-.02 (-2.83)	-.005(-.24)	-.02 (-1.04)
16 Class Size ≥ 33	-.16 (-.10)	-.82 (-.54)	.13 (.08)	.20 (.13)
17. Class Size, 28–33	-4.22 (-1.74)	-.20 (-.14)	-1.32 (-.52)	.17 (.17)
18 (17) x Third-Grade Score	.14 (2 22)	-	.06 (.79)	-
19. School Enrollment	-.002(-1.54)	-.003(-2.33)	-.001(-.33)	-.003(-1.80)
20 (19) x Race	-.004(-1.65)	-	-.005(-1.84)	-
21 Library Books/Pupil	-.43 (-1.97)	-.47 (-2.21)	-.12 (-.70)	-.13 (-.81)
22. Percent Blacks ≥ 20 and < 40	3.70 (2.53)	3.08 (2.15)	3.99 (2.49)	3.81 (2.47)
23 Percent Blacks ≥ 40 and < 60	6.02 (3.61)	5.64 (3.34)	4.02 (2.37)	3.55 (2.13)
24. Percent Blacks ≥ 60	4.18 (2.62)	3.10 (1.93)	3.95 (2.34)	2.74 (1.69)
25. Percent High Achievers	.68 (2.55)	.15 (1.60)	.69 (2.05)	.17 (1.75)
26. (25) x Third-Grade Score	-.01 (-2.20)	-	-.01 (-1.73)	-
27 Percent Low Achievers	-.08 (-2.11)	-.08 (-2.28)	-.07 (-1.52)	-.08 (-1.85)
28. Disruptive Incidents	1.86 (3.89)	.35 (2.09)	1.50 (2.69)	.37 (2.23)
29. (28) x Third-Grade Score	-.05 (-3.18)	-	-.03 (-2.00)	-
Constant	22.09	25.09	11.70	21.82
$\bar{R}^2$	.27	.24	.23	.21
$\hat{F}$	9.03	10.68	7.50	9.52

Note: Dependent Variable: Sixth-Grade Score on Iowa Test of Basic Skills (Composite) Minus Third-Grade Score.

[a]Because they are less pupil-specific, variables 11 through 29 are not identical in columns (4) and (5) to those in columns (2) and (3). Exact definitions for all the variables are in the Appendix.

[b]When school averages were tested, it was no longer possible to use the best fit breaking point of the 525 undergraduate college rating, because too few averaged observations were at or above 525. A rating of 450 was used, therefore. In column (2) the intercept term had a coefficient of 4.23 and a t-statistic of 1.82; the interaction term had a coefficient of −.49 and a t-statistic of −1.95. In column (3), the term reads −.10(−.14).

[c]Sources of data are listed in the Appendix.

grew less. Interestingly, it was the middle and higher income students whose growth in scores registered the greatest declines when their Unexcused Absences were greater—perhaps, because when an advantaged student misses school it signifies a far more serious negative attitude toward schooling than when a disadvantaged student does the same thing. (Five more unexcused absences were associated with 2.1 months less growth over three years for students with a $10,000 family income, and only .7 months less for those with a $5,000 income.) In the case of Latenesses, however, it was only the lower income students where the effect was pronounced.

Other socioeconomic factors examined, but not found to be significant, were the number of residential moves of the students, whether or not the pupil was foreign born, and whether or not both parents were foreign born.

B. *School Inputs: Teacher Quality*[13]

Finding out which school inputs are helpful to learning is what we are, of course, most interested in. These are the things that school administrations and teachers' unions can, if the spirit is willing, do something about.

The unique detail of the data base on the characteristics of teachers makes possible a deeper inquiry into the question "what matters?" than is possible with noisy, aggregative data. Several "quality" measures of each of the teachers were obtained and the specific teachers each pupil had were identified. What mattered?[14] Teachers who received B.A.'s from higher rated colleges[15] (rows 11 and 12) were associated with students whose learning rate was greater—and it was students from lower income families who benefitted most. (A student whose family income was $5,000 grew 8.6 months more with a teacher from a higher rated college than with teachers from other colleges; a student whose family income was $10,000 grew 3.7 months more.) Teacher experience (rows 13 and 14) has been found to be unimportant in many studies.[16] Though it seems clearly unreasonable to expect equal effectiveness of experience for all levels of student abilities, the interaction between them has not been fully examined. We found that students whose third-grade score was above grade level benefitted from more experience, but those who were very much below grade level were negatively affected. In fact, this latter group did best with newer teachers who perhaps have an undampened enthusiasm for teaching those who find it hard to learn. These teacher quality findings were robust. They remained essentially intact whether all or some of the teacher quality variables were controlled, and when the sample was broken into subsamples by

[13]The qualities of sixth-grade teachers, rather than fourth-, fifth-, and sixth-grade teachers, were used. Evidence that the sixth-grade teachers had the strongest impact piled up: The $\bar{R}^2$ for the equation using sixth-grade teachers was significantly higher than using either of the other years. The Koych lag formulation (see Zvi Griliches) which assumes that the weights assigned diminish geometrically for more distant periods was used. The $\bar{R}^2$'s, coefficients, and t-statistics for the college rating and experience of teachers, for the equation which gives most weight to the sixth-grade teacher qualities, and least to the fourth and fifth, are significantly higher than in the alternate specifications. They diminish continuously as the weight assigned to the sixth-grade teacher decreases.

[14]Hanushek (1970) and Murnane both had some teacher data which were specific to the pupil. Both used dummy variables for the teachers, testing to find out whether or not teachers mattered. For his subsample of Anglo-American students, Hanushek showed the quality of teachers mattered, but not for the subsample of Mexican-Americans (perhaps, he hypothesized, because of the language problem). Murnane concluded that teaching quality did matter. While the use of teacher dummies did not demonstrate that teacher quality can be proxied by any objectively measurable quality, it does give support to the conclusion that the use of pupil-specific teacher variables results in more revealed impact on achievement.

[15]Winkler, in California, also found that graduates of "good" colleges taught significantly better than graduates of "bad" colleges.

[16]Coleman found that teacher experience mattered only marginally. Hanushek found no statistical significance to teacher experience in any of his three California subsamples. Murnane found that having experience of one, two, or three years over having none mattered greatly; the extra benefit of between three and five years was less; and, beyond five years, he found no effect of experience on achievement.

TABLE 2—RESULTS ON ACHIEVEMENT OF TEACHER'S COLLEGE RATING AND EXPERIENCE USING SUBSAMPLES

	Low Income[a]	Middle Income[a]	High Income[a]
Rating of Teacher's College, I ≥ 525[b]	15.32 (4.16)	.09 (.03)	3.49 (1.64)
	Low Achievers[c]	Middle Achievers[c]	High Achievers[c]
Teacher's Experience[d]	−.01 (−.04)	.16 (.91)	.31 (1.64)

[a]Low income is defined as < $7000, middle income as ≥ $7000 and ≤ $9000, and high income as > $9000.

[b]The comparable coefficients and tests of significance, using interaction terms, were 8 62 (2.95) for students whose family income was $5000, 5.64 (3 28) for those at an $8000 income level, and −1 29 (−0.36) for those at a $15,000 level

[c]Low achievers are those whose third-grade Iowa Test scores were in grade equivalent terms ≤ 2.7; middle achievers scored > 2.7 and < 3 9, high achievers scored ≥ 3 9.

[d]The comparable coefficients and tests of significance, using interaction terms, were − 29 (−1.54) for students who scored two grades below level (at 1 0), .05 (0.53) for those who were at grade level (3.0), and .39 (2 42) for those who were two grades above (at 5 0)

race, achievement, and income. The relevant coefficients and t-statistics for the subsamples are shown in Table 2.

Finally, we found a perverse relationship between the National Teacher Exam Score (row 15) and learning. The discriminatory powers of the exam were evaluated by the Philadelphia School District in 1972. They concluded that the scores should not be used as the only measure of the potential of a teacher—these findings suggest that they should not be used as any measure.

C. *School Inputs: Nonteacher School Quality*

Does Class Size (rows 16, 17, and 18) matter? Educators' studies cover the full range of answers. One summary report (see Howard Blake) stated that thirty-five of the studies surveyed found that smaller classes were more effective, eighteen found that larger classes were more effective, and thirty-two were inconclusive! More light is shed, however, when the differential effects of smaller and larger classes are examined in relation to achievement levels (rows 16, 17, 18). Low-achieving students (those scoring two or more years below grade level) did worse (almost three months worse over the three years) in classes with more than 28 students; high-achieving students (those scoring two or more grades above grade level) did better (almost three months better); those around grade level appeared to be unaffected. When the Class Size variable was explored through ability subsamples, rather than interactions, the same results emerged. The relevant coefficients and t-statistics are shown in Table 3.

School Enrollment (rows 19 and 20), also a variable about which studies have produced a wide range of possible impact, is shown in this study to have differential effects on black and nonblack students. Smaller schools appeared to be better for all, but had a larger beneficial effect on achievement growth for black pupils. Library Books per pupil (row 21) had a perverse effect—possibly because the fact that libraries were larger in relation to the number of pupils in schools where there were more low-income pupils was not adequately picked up by the other variables.[17]

There appear then to be a number of school inputs that can raise student achievement. Equally interesting was the discovery that a number of inputs which we pay for did not accomplish this goal. 1) The general physical facilities of schools did not seem to make much difference, one way or another, to students' learning. Whether the pupil

[17]Coleman stated that "The number of volumes per student in the school library shows small and inconsistent relations to achievement" (p. 316).

TABLE 3—RESULTS FOR CLASS SIZE USING ABILITY-GROUPED SUBSAMPLES[a]

	Low Achievers	Middle Achievers	High Achievers
Class Size ≥ 33[b]	−2.75 (−1.22)	1.48 (.58)	3.00 (.75)
Class Size, 28–33[b]	−4.27 (−1.98)	2.12 (.89)	5.42 (1.43)

[a]For definitions of subsamples, see fn. c, Table 2

[b]The comparable coefficients and tests of significance, using interaction terms, were −2.83 (−1.44) for students who scored two grades below level (at 1.0), −.05 (−.04) for those who were at grade level (3.0), and 2.72 (1.48) for those who were two grades above (at 5.0).

had access to more or less playground space, a new or old building, or a building rated higher or lower in general physical condition, did not seem to matter much when it comes to achievement test scores. 2) There was a wide range among Philadelphia elementary school principals in experience, extra degrees, and extra educational credits—but there were no discernible differences in the impact on achievement. 3) Whether teachers have more or less education beyond the B.A. did not seem to make his or her students learn more. The finding of an absence of impact on achievement of extra training is consistent with many studies[18]—yet teacher salary scales nearly everywhere incorporate substantial rewards to teachers who take additional educational credit beyond the B.A.

Clearly, though these inputs do not seem to relate to achievement growth, it does not mean that reducing these expenditures to zero is the logical policy recommendation. Without some minimum level, there might be some negative effect, and with much more than is now put in, there might be some positive effect. Further, and a point not to be ignored, the objective of many school inputs is not limited to the one used here—growth in achievement scores. Other objectives—attitudes toward other races, sense of participation in the democratic process—may be part of the desired outcome.

[18]Coleman, Burkhead et al., Hanushek, Kiesling, and Winkler all found that having or not having an M.A. was not significantly related to achievement. The literature in education journals on this subject reveals results on both sides.

D. *Peer Group Effects*

School segregation, in the sense of heavily concentrated racial mixtures, exists in the Philadelphia schools. In the years covered by the study sample, 23 percent of the elementary schools had less than 10 percent black students, and 40 percent had more than 90 percent. Busing for desegregation did not exist, and could obviously introduce entirely new elements not reflected in our findings. We found that black and nonblack students benefitted (rows 22, 23, and 24)—had the largest growth in achievement—when they were in schools with a 40 to 60 Percent Black student body, rather than in schools that were more racially segregated.[19] From a policy point of view, of course, what one is interested in is the effect of changes in the proportion of black students when variables associated with these changes are not controlled. The finding that the largest growth in achievement took place with a 40 to 60 Percent Black student body remained intact, when the variables measuring the Proportion of High

[19]One is tempted to dissect this somewhat unexpected (for nonblacks) result by searching for characteristics of the families that live in racially mixed areas that would contribute to learning growth, but which are not captured by the variables in the equation. In addition to the family income variable already in, variables measuring the average income of residents in the school feeder area, the change in relative income of the feeder area between the 1960 and 1970 Censuses, and the average education level of the 25-year and over population were tried and found to be entirely without significance. In addition, dummy variables for the two areas of Philadelphia which have had the most stable pattern of racial integration were explored and found to be not significant.

TABLE 4—RESULTS ON ACHIEVEMENT GAIN OF THE PROPORTION OF BLACK STUDENTS WITH AND WITHOUT CONTROLLING FOR ASSOCIATED VARIABLES

Independent Variable	Controlling for Variables 25, 26, 27, 28, 29	Without Controlling for Variables 25, 26, 27, 28, 29
22. Percent Blacks ≥20 and <40	3.70 (2.53)	2.45 (1.73)
23. Percent Blacks ≥40 and <60	6.02 (3.61)	4.25 (2.76)
24. Percent Blacks ≥60	4.18 (2.62)	2.80 (2.04)

Achievers (rows 25 and 26), the Proportion of Low Achievers (row 27), and the number of Disruptive Incidents (rows 28 and 29) were not included. (See Table 4.) The coefficients and t-statistics were slightly lower.

Ability grouping can be regarded as another type of segregation. What are the effects of being in a grade with more or less very high or very low achievers (rows 25, 26, and 27)? Seventy percent of the elementary schools had less than 10 percent high achievers; 52 percent had more than 50 percent of the student body achieving at very low levels. In elementary schools, students who test at grade level or lower are distinctly helped by being in a school with more high-achieving students. The students who are performing above their grade level are not particularly affected—it is the low-achieving students who benefit most. This finding is so, whether or not the proportion of black students is controlled. Being in a student body with more low achievers has a negative effect on learning for all students, though this is not so if there is no control for racial composition. What all this seems to say is that high achievers are relatively unaffected by variations in the percentage of top achievers. But, for the low achievers, the intellectual composition associated with other characteristics of the student body has a direct impact on learning.

Finally, what is one to say about the finding that, for students who are at or below grade level, more Disruptive Incidents (rows 28 and 29) are associated with greater achievement growth? It is only for those well above grade level that we found the expected negative effect. This may well be a case where very different results would be obtained if the data were available on a classroom-specific basis. If disruptive incidents only broke out in classrooms where achievement growth barely occurred, then spreading the impact among all classes may have produced this anomalous effect. In any case, it would seem a bit premature to engage in a policy of encouraging disruptive incidents to increase learning!

Again, it is interesting to note a number of peer group characteristics that were not significant in any of the specifications: percent of the student body receiving free lunches, student mobility in the school, median income of the school feeder area, average education level of adults, (25 years and over) in the school feeder area, change in the relative income of the area from 1960 to 1970, and average daily attendance in the school.

In summary, then, we find that many school inputs mattered, and that their impact varied considerably on different types of students. There were inputs that helped disadvantaged students. Clearly, this is consistent with what educators and parents believe when they advocate individualizing education.

IV. What Was Revealed By Interaction Analysis of Inputs?

On the basis of this study, at least, it would appear that there is much to be gained from examining the differential impact (by race, income, and achievement level) of educational inputs on achievement growth. How much there is to be gained can be seen more precisely by a comparison of columns (2) and (3) in Table 1. In column (3), the results that would have been

produced by this pupil-specific data base, if interactions had not been analyzed, are given. Apart from the reduced explanatory power of the equation, three general types of insight emerge.

First, the expected negative effects of low income and race have been so extensively traced through specific inputs in column (2) that the Income and Race intercepts have no significance there at all. But when they are not traced through, in column (3), the coefficient of the Income variable becomes larger and the Race variable acquires a very large and significant negative coefficient.[20] This is, of course, consistent with the important and repeated finding of many studies searching for the determinants of student achievement. But, the lack of insight into just where—in connection with which inputs—Income and Race play a role, has left schools relatively impotent in changing the achievement growth of students.

Second, a possible explanation for the hopelessness of the findings of many studies is that the effectiveness of some inputs may be of directly opposite direction for different types of students. We found this in the case of three variables. Without interaction analysis, the conclusions would have been that there was no additional positive effect of having an IQ of 110 or more, that the experience of teachers had no effect on achievement growth, and that class size was irrelevant to learning (rows 5, 13, 16, and 17, column (3)). Interaction analysis revealed that the nonblack student with an IQ of 110 or more moved much more rapidly ahead, but the black student did not; that higher achieving students benefitted by being with more experienced teachers,[21] but lower achieving students were adversely affected—in "fact," benefitting from newer teachers; and that low achievers did better in classes with less than 28 students than in classes with 28 to 33, while high achievers did worse.[22]

Finally, without examining the interactive effects of inputs, the policy implications of important differential effects (not necessarily of opposite sign) are lost. From column (3) it would not be possible to "know" that higher income students were most negatively affected by more Unexcused Absences,[23] that only the low-income students were adversely affected by more latenesses, that the lower income students benefitted most (and considerably) by having teachers who came from higher rated colleges, that black students experienced the bigger boost to learning from being in smaller schools (School Enrollment), that the low achiever is the real beneficiary of being in a grade with higher Percent High Achievers (the high achiever is unaffected), and that the expected negative effect of being in a school with more serious Disruptive Incidents shows up only for the very high achievers.

V. What Was Revealed By Pupil-Specific Data on School Inputs?

It would appear from comparing the results using pupil-specific data on inputs (Table 1, column (2)) with the results using school averages of inputs (column (4)) that there is much to be learned from the use of the more disaggregated data base. Many studies have been hampered by the limited amount of data available (easily obtainable?) which are specifically tied to the pupil. Thus, in Coleman's seminal work, and in the reanalysis of his *EEO* data, the only data used bearing on the influence of teacher experience were the average experience levels in schools—not, as we were able to obtain for this study of Philadelphia

[20]Results, similar in direction, showed up in the eighth-grade sample.

[21]In the eighth-grade sample, the experience of the social studies teacher lost significance when interactions were not explored.

[22]The negative impact of being in a class below thirty-two was reduced, and the very large negative effect for low income students of being in classes larger than thirty-two was lost, when interactions were not used in the eighth-grade sample

[23]In the eighth-grade sample, it would not have been possible to "know" that average and above average achievers were the only ones negatively affected.

schools, the experience level of the specific teachers confronting each pupil.

The most interesting results relate to the totally changed findings on the teacher variables. The rating of their undergraduate colleges (Table 1, rows 11 and 12), their experience (rows 13 and 14), and their teacher exam score (row 15) were not significant when averages were used. And, equally interesting, the class size (rows 16, 17, and 18), showed no impact when it was measured in a way not specific to the pupil.[24]

Other school input and climate variables which, in column (2), were changed over the three years the student was tracked (so that if the student changed schools, a change in the school variable was incorporated) were reduced in significance, when the tracking was not done. Thus, when just the sixth-grade value of the variable was used, the impact of the number of library books (row 21), of the size of the school on nonblacks, (rows 19 and 20), of the proportions of high- and low-achievers (rows 25, 26, and 27), and the proportions of black students (rows 22, 23, and 24) in the school were reduced.

VI. Concluding Remarks

The findings of this study suggest 1) that when there are extensive pupil-specific data available, more impact from school inputs is revealed; 2) that when the effects of school inputs are examined differentially, more impact and insight emerge; 3) that most of the effects of family income and race can be tagged to specific school inputs; and 4) that the low achiever, the low-income student, and the black student do respond in terms of achievement growth to some school inputs.

[24]In the eighth-grade sample, when the same comparison was made between the results using pupil-specific data and not, similar findings emerged: the impact of the undergraduate college rating for social studies teachers became weaker, the experience of English teachers lost significance, the experience of math teachers was reduced in significance, and class size had no significance.

A comparison of the results in column (5) (where school-wide averages of school inputs were the data and no interaction analysis was used) with the results in column (2) (where the data on school inputs were specific to the pupil and differential impacts were examined) should encourage the pursuit of disaggregated data bases. No impact from teacher experience or the exam score of the teacher emerges (rows 13 and 15). The rating of the teacher's college has a perverse effect (row 11). Class size has no effect (rows 16, 17, and 18). And the impact of the proportion of blacks (rows 22, 23, and 24), the proportion of high achievers (rows 25 and 26), and the proportion of low achievers (row 27) is reduced.

Perhaps most interesting from the point of view of those seeking to make more equal the achievement growth of the advantaged and the disadvantaged, is the finding that there appear to be school inputs which help the disadvantaged do better. If the courts are to rule on these matters, then targeting which school resources are specifically helpful to the disadvantaged is essential. If, for example, physical facilities are not particularly helpful to achievement, then learning that schools in better condition tend to be in locations where there are fewer disadvantaged students should not necessarily lead to action to equalize the facilities. If, on the other hand, the fact that smaller classes are particularly helpful to the learning growth of low achievers is verified, then there is something specific to advocate in the quest for equity. The findings of this study suggest that there are school inputs which help the low achievers to do better. And most of these were findings which emerged only when pupil-specific data were used and examined in relation to student characteristics.

APPENDIX

The definitions, $\bar{X}$'s, and σ's of each of the variables shown in Table I are listed here with the corresponding row numbers. The prime notation is used for the variables

in columns (4) and (5) which correspond to the variables in columns (2) and (3).

1. Income: Estimated family income (in thousands); $\bar{X}$ = \$8.802; σ = \$2.609.
2. Race: Dummy variable, 0 = nonblack, 1 = black; $\bar{X}$ = .64; σ = .48.
3. Sex: Dummy variable, 0 = female, 1 = male; $\bar{X}$ = .48; σ = .50.
4. IQ: Score on first-grade Philadelphia Verbal Ability Test; $\bar{X}$ = 100.65; σ = 14.46.
5. IQ 110+: Additional effect of scoring 110 or more on IQ test; form is a two-piece linear function with corner at 110.
6. (5) x Race: Interaction of Race with IQ 110+.
7. Unexcused Absences: Average of annual number of unexcused absences over three years, 1967/68–1970/71; $\bar{X}$ = 2.33; σ = 4.48.
8. (7) x Income: Interaction of Income with Unexcused Absences.
9. Latenesses: Average of annual number of latenesses over three years, 1967/68–1970/71; $\bar{X}$ = 2.02; σ = 3.59.
10. (9) x Income: Interaction of Income with latenesses.
11. Rating of Teacher's College, 1 ≧ 525: Gourman rating of sixth-grade teacher's undergraduate college, 0 = <525, 1 = ≧525; $\bar{X}$ = .05; σ = .22.

11′. School $\bar{X}$ of Ratings of Teacher's Colleges, 1 ≧ 450: School average of Gourman ratings of all teachers' undergraduate colleges, 0 = <450, 1 = ≧450; $\bar{X}$ = .04; σ = .21.

12. (11) x Income: Interaction of Income with Rating of Teacher's College, 1 ≧ 525.

12′. (11′) x Income: Interaction of Income with School $\bar{X}$ of Ratings of Teacher's Colleges, 1 ≧ 450.

13. Teacher's Experience: Sixth-grade teacher's experience (in years up to 11); $\bar{X}$ = 6.58; σ = 3.71.

13′. School $\bar{X}$ of Teacher's Experience: School average of all teachers' experience (in years up to 11); $\bar{X}$ = 7.28; σ = 1.08.

14. (13) x Third-Grade Score: Interaction of Third-Grade Score with Teacher's Experience.

14′. (13′) x Third-Grade Score: Interaction of Third-Grade Score with School $\bar{X}$ of Teacher's Experience.

15. National Teacher Exam Score: Sixth-grade teacher's score on National Teacher Examination, Common; $\bar{X}$ = 606.83; σ = 58.93.

15′. School $\bar{X}$ of Teacher's National Teacher Exam Scores: School average of all teachers' National Teacher Examination, Common, scores; $\bar{X}$ = 606.91; σ = 19.94.

16. Class Size ≧ 33: Dummy variable, 1 = ≧33, 0 = <33; class size is the average size of the class in pupil's grade in pupil's school, $\bar{X}$ = .30; σ = .46.

16′. School $\bar{X}$ of Class Size ≧ 33; Dummy variable, 1 = ≧33, 0 = <33; class size is average of all classes in school; $\bar{X}$ = .17; σ = .38.

17. Class Size, 28–33; Dummy variable, 1 = ≧28 and <33, 0 = <28 or ≥33; class size is the average size of the class in pupil's grade in pupil's school; $\bar{X}$ = .63; σ = .48.

17′. School $\bar{X}$ of Class Size, 28–33; Dummy variable, 1 = ≧28 and <33, 0 = <28 or ≥33; class size is average of all classes in school; $\bar{X}$ = .66; σ = .48.

18. (17) x Third-Grade Score: Interaction of Third-Grade Score with Class Size, 28–33.

18′. (17′) x Third-Grade Score: Interaction of Third-Grade Score with School $\bar{X}$ of Class Size, 28–33.

19. School Enrollment ($\bar{X}$): Average of number of pupils enrolled in each school pupil attended, 1967/68–1970/71; $\bar{X}$ = 917.35; σ = 336.05.

19′. School Enrollment (6): Number of pupils enrolled in school pupil attended in sixth grade; $\bar{X}$ = 912.90; σ = 343.79.

20. (19) x Race: Interaction of Race with School Enrollment ($\bar{X}$).

20′. (19′) x Race: Interaction of Race with School Enrollment (6).

21. Library Books/Pupil ($\bar{X}$): Average number of library books per pupil in each school pupil attended, 1967/68–1970/71; $\bar{X}$ = 6.98; σ = 2.37.
21′. Library Books/Pupil (6): Number of library books per pupil in school pupil attended in sixth grade; $\bar{X}$ = 7.88; σ = 3.00.
22. Percent Blacks $\geqq$ 20 and < 40 ($\bar{X}$): Dummy variable: 1 = Percent blacks in school $\geqq$ 20% and < 40%; average of each school pupil attended, 1967/68–1970/71; $\bar{X}$ = .09; σ = .29.
22′. Percent Blacks $\geqq$ 20 and < 40 (6): Dummy variable described in 22, for sixth-grade school; $\bar{X}$ = .08; σ = .27.
23. Percent Blacks $\geqq$ 40 and < 60 ($\bar{X}$): Dummy variable: 1 = Percent blacks in school $\geqq$ 40% and < 60%; average of each school pupil attended, 1967/68 1970/71; $\bar{X}$ = .10; σ = .31.
23′. Percent Blacks $\geqq$ 40 and < 60 (6): Dummy variable described in 23, for sixth-grade school; $\bar{X}$ = .11; σ = .32.
24. Percent Blacks $\geqq$ 60 ($\bar{X}$): Dummy variable: 1 = Percent blacks in school $\geqq$ 60%; average of each school pupil attended, 1967/68-1970/71, $\bar{X}$ = .52; σ = .50.
24′. Percent Blacks $\geqq$ 60 (6): Dummy variable described in 24, for sixth-grade school; $\bar{X}$ = .51; σ = .50.
25. Percent High Achievers: Average percent in pupil's fifth and sixth grades scoring above 84th National Percentile on Iowa Test of Basic Skills; $\bar{X}$ = 4.00; σ = 5.93.
25′. Percent High Achievers in School: Percent of pupils in all grades of pupil's sixth-grade school scoring above 84th National Percentile on Iowa Test of Basic Skills; $\bar{X}$ = 5.56, σ = 6.55.
26. (25) x Third-Grade Score: Interaction of Third-Grade Score with Percent High Achievers.
26′. (25′) x Third-Grade Score: Interaction of Third-Grade Score with Percent High Achievers in School.
27. Percent Low Achievers: Average Percent in pupil's fifth and sixth grades scoring below 16th National Percentile on Iowa Test of Basic Skills; $\bar{X}$ = 48.75; σ = 18.50.
27′. Percent Low Achievers in School: Percent of pupils in all grades of pupil's sixth-grade school scoring below 16th National Percentile on Iowa Test of Basic Skills; $\bar{X}$ = 45.41; σ = 17.39.
28. Disruptive Incidents ($\bar{X}$): Average of annual number of disruptive incidents in each school pupil attended, 1967/68–1970/71; $\bar{X}$ = 2.89; σ = 2.31.
28′. Disruptive Incidents (6): Number of disruptive incidents in school and year pupil attended in sixth grade; $\bar{X}$ = 2.85; σ = 2.39.
29. (28) x Third-Grade Score: Interaction of Third Grade Score with Disruptive Incidents ($\bar{X}$).
29′. (28′) x Third-Grade Score: Interaction of Third Grade Score with Disruptive Incidents (6).

Sources: 1: The authors (1975), Individual Pupil Records, Form EH-7, and School District of Philadelphia (*SDP*) 1960–70 Tape on Pupil Addresses. 2–10: Individual Pupil Records, Form EH-7 and Roll Sheets. 11 12′: Jack Gourman and *SDP* Permis File. 13–15′: *SDP* Permis File. 16–20′: Monthly Reports, Field Operations and Subsidies Division, *SDP*. 21–21′: Office of Research and Evaluation, *SDP*. 22–24′: *Enrollment, Negro and Spanish-Speaking in the Philadelphia Public Schools.* 25–27: *Reports of Division of Testing*, Office of Research and Evaluation, *SDP*. 25′–27′: Office of Research and Evaluation, *Reports of Spring Achievement Test Results.* 29–29′: Facilities Security, *SDP*, *Reports on Incidents by School Code, Date, and Incident Classifications.*

REFERENCES

H. A. Averch et al., "How Effective is Schooling? A Critical Review and Synthesis of Research Findings," The Rand Corporation, R-956-PCSF/RC, Santa Monica 1972.

H. A. Averch and H. J. Kiesling, "The Relationship of School and Environment to Student Performance: Some Simultaneous Models for the Project TALENT High Schools," The Rand Corporation, mimeo., Santa Monica 1970.

Charles Benson et al., *State and Local Fiscal Relationships in Public Education in California*, California State Senate Fact Finding Committee on Revenue and Taxation, Sacramento 1965.

H. Blake, "Class Size: A Summary of Selected Studies in Elementary and Secondary Public Schools," unpublished doctoral dissertation, Columbia Univ. 1954.

S. Bowles, "Toward an Educational Production Function," in W. Lee Hansen, ed., *Education, Income, and Human Capital*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 35, New York 1970, 11–60.

Jesse Burkhead et al., *Input and Output in Large City High Schools*, Syracuse 1967.

E. Cohn, "Economies of Scale in Iowa High School Operations," *J. Hum. Resources*, Fall 1968, *3*, 422–34.

James S. Coleman et al., *Equality of Educational Opportunity*, Washington 1966.

J. S. Coleman and N. L. Kaweit, *Measures of School Performance*, Santa Monica 1970.

Jack Gourman, *The Gourman Report*, Phoenix 1967.

Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.

Eric Hanushek, *Education and Race*, Lexington 1972.

———, The Production of Education, Teacher Quality, and Efficiency," in *Do Teachers Make a Difference?*, Office of Education, Department of Health, Education and Welfare, Bureau of Educational Personnel Development, Washington 1970, 79–99.

Christopher Jencks et al., *Inequality: A Reassessment of the Effect of Family and Schooling in America*, New York 1972.

Martin T. Katzman, *The Political Economy of Urban Schools*, Cambridge 1971.

Herbert J. Kiesling, *A Study of Cost and Quality of New York School Districts*, Washington 1970.

H. M. Levin, "A New Model of School Effectiveness," in *Do Teachers Make a Difference?*, Office of Education, Department of Health, Education and Welfare, Bureau of Educational Personnel Development, Washington 1970, 55–78.

S. Michelson, "The Association of Teacher Resourceness with Children's Characteristics," in *Do Teachers Make a Difference?*, Office of Education, Department of Health, Education and Welfare, Washington 1970, 120–68.

Richard J. Murnane, *The Impact of School Resources on the Learning of Inner City Children*, Cambridge 1975.

Frederick Mosteller and Daniel P. Moynihan, *On Equality of Educational Opportunity*, New York 1972.

A. Summers and B. Wolfe, "Equality of Educational Opportunity Quantified: A Production Function Approach," *Philadelphia Fed Research Papers*, 1974.

—— **and** ——, "Intradistrict Distribution of School Inputs to the Disadvantaged: Evidence for the Courts," *J Hum. Resources*, Summer 1976, *11*, 328–42.

—— **and** ——, "Manual on Procedure for Using Census Data to Estimate Block Income," in *Philadelphia Fed Research Papers*, 1975.

D. R. Winkler, "Educational Achievement and School Peer Composition," *J. Hum Resources*, Spring 1975, *10*, 189–204.

School District of Philadelphia (*SDP*), "Evaluation of National Teacher Examination Research Study," *Final Report of the Task Force on Teacher Selection in Philadelphia*, Philadelphia 1972.

The Forward Exchange Rate, Expectations, and the Demand for Money: The German Hyperinflation

By JACOB A. FRENKEL*

A major difficulty in incorporating the role of inflationary expectations in empirical work has been the lack of an observable variable measuring expectations. Thus, for example, in analyzing the demand for money during hyperinflation, Phillip Cagan in his classic contribution constructed a time-series of expected inflation using a specific transformation of the time-series of the actual rates of inflation. There are two conceptual difficulties with such an approach: first, the choice of the specific transformation used to generate the series of expectations is to a large extent arbitrary; and second it assumes that expectations about future prices are based only on past and present prices. Recent empirical work that was stimulated by Cagan's pioneering study elaborated on some aspects of the estimation procedures (see Thomas Sargent and Neil Wallace; Sargent 1977; Joseph Bisignano; Paul Evans; Rodney Jacobs 1975; Mohsin Khan 1975), and the functional form (see Robert Barro 1970; Benjamin Eden 1976). The ongrowing literature concerning rational expectations (for example, John Muth 1961; Robert E. Lucas) has led to an examination of the conditions under which the adaptive expectations process is "rational" in the sense of Muth (1961). See Sargent and Wallace (1973); Sargent 1977); Benjamin Friedman (1975a); Michael Mussa.

In this paper I propose a direct measure of expectations which is then incorporated in the analysis of the demand for money during the German hyperinflation. The major virtue of the proposed direct measure is that it is not derived from a specific mechanistic formula, but rather, it reflects the expectations of economic agents as manifested in market prices. The direct measure is based on data from the forward market for foreign exchange. The plan of the paper is as follows: Section I describes the direct measure of expectations and provides evidence on the efficiency of the foreign exchange market. Section II incorporates these expectations in estimating the demand for money. The issues that are discussed in that section involve the proper functional form, the proper price deflator, the stability of the demand for money during the various phases of the hyperinflation, possible lags of adjustment and the resultant estimates of short-run and long-run demand functions, and the role of price variability and uncertainty in the specification of the demand for money. Section III deals with the issue of inflationary finance and the money supply process. In this context I examine the interrelationships between money and prices and discuss some aspects of "causality" by analyzing the time-series properties of money and prices. Section IV contains some concluding remarks.

*University of Chicago and Tel-Aviv University. I am indebted to John Bilson and Rolf Banz for suggestions and efficient research assistance. In revising the paper I have benefited from numerous suggestions by Robert Barro, Phillip Cagan, Kenneth Clements, Rudiger Dornbusch, Paul Evans, Stanley Fischer, Benjamin Friedman, Milton Friedman, John Gould, Zvi Griliches, Arnold Harberger, Albert Hart, James Heckman, Edi Karni, Mohsin Khan, David Laidler, Edward Lazear, Robert Lucas, Huston McCulloch, Merton Miller, Franco Modigliani, Michael Parkin, Aris Protopapadakis, Thomas Sargent, Jose Scheinkman, Larry Sjaastad, Jerome Stein, Lester Telser, and Arnold Zellner. Financial support was provided by a grant from the Ford Foundation.

I. The Foreign Exchange Market and Expectations[1]

A. *Expectations and the Forward Exchange Rate*

The fundamental variable that is used in representing the market measure of expectations is derived from the market for foreign exchange. In that market, the premium (or discount) on a forward contract for foreign exchange measures the anticipated depreciation (or appreciation) of the domestic currency in terms of foreign exchange. To the extent that domestic money is held as a substitute for foreign money, the demand for domestic money will depend on the anticipated change in the exchange rate as measured by the forward premium (or discount). To the extent that domestic money is held as a substitute for bonds, we may still use the forward premium as measuring the relevant cost of holding money by relying on the interest parity theory. That theory maintains that in equilibrium the premium (or discount) on a forward contract for foreign exchange is (approximately) related to the interest rate differential according to:

$$\frac{F - S}{S} = i - i^* \tag{1}$$

where F and S are the forward and spot exchange rates (the domestic currency price of foreign exchange), respectively; i the domestic rate of interest; and i^* the foreign rate of interest on comparable securities for the same maturity. Evidence available for various countries over various time periods suggest that this parity condition holds (see the author and Richard M. Levich 1975, 1977), although, due to lack of data, no comparable study has been done on the period of the German hyperinflation. It is reasonable to assume that during the hyperinflation most of the variations in the difference between domestic and foreign anticipated rates of inflation were due to anticipated domestic (German) inflation. It follows, therefore, that the variations of the forward premium on foreign exchange $(F - S)/S$ may be viewed as a measure of the variations in the expected rate of inflation (as well as the expected rate of change of the exchange rate).

[1]The argument in this section draws on my paper (1976a).

B. *The Efficiency of the Foreign Exchange Market*

Prior to incorporating the forward premium as a measure of expectations, it is pertinent to explore the efficiency of the foreign exchange market during the turbulent hyperinflation period. Evidence on the efficiency of that market will support the approach of using data from that market as the basis for inference on expectations.

If the foreign exchange market is efficient and if the exchange rate is determined in a similar fashion to other asset prices, we should expect the behavior in that market to display similar characteristics as those displayed in other stock markets. In particular, we should expect that current prices reflect all available information, and that the residuals from the estimated regression should be serially uncorrelated.

To examine the efficiency of the market we first regress the logarithm of the current spot exchange rate, $\log S_t$, on the logarithm of the one-month forward exchange rate prevailing at the previous month, $\log F_{t-1}$.

$$\log S_t = a + b \log F_{t-1} + u \tag{2}$$

The expectation is that the constant term does not differ significantly from zero, that the slope coefficient does not differ significantly from unity, and that the error term is serially uncorrelated.[2] Since data on the German Mark-Pound Sterling (RM/£)

[2]The logarithms of the exchange rate are used rather than the exchange rate itself. In the latter case the regression is completely dominated by the last few observations due to the rapid acceleration of the depreciation of the Reichsmark (RM) Furthermore, results obtained by using changes in the logarithms may be interpreted in terms of a comparison between expected and realized rates of return. Data on the spot and forward exchange rates are from Paul Einzig.

forward exchange rate are available only from February 1921, equation (2) was estimated over the period February 1921–August 1923 (31 months). The resulting ordinary least squares estimates are reported in equation (2′) with standard errors in parentheses below the coefficients.

$$(2') \quad \log S_t = \underset{(.25)}{-.45} + \underset{(.03)}{1.09} \log F_{t-1}$$
$$\bar{R}^2 = .98; s.e. = .46; D.W. = 1.89$$

As can be seen, the constant term does not differ significantly from zero at the 95 percent confidence level (although it seems to be somewhat negative), the slope coefficient is somewhat above unity (at the 95 percent confidence level) but, most importantly, the Durbin-Watson (*D.W.*) statistic indicates that the residuals are not serially correlated.[3] The fact that the slope coefficient is slightly above unity may be explained in terms of transaction costs or in terms of the Keynesian concept of normal-backwardation (see John M. Keynes, 1930 Vol. II, p. 143).[4] The joint hypothesis that the constant term is zero *and* that the slope coefficient is unity is rejected at the 95 percent confidence level.

To explore further the implications of the efficient market hypothesis we examine whether the forward exchange rate summarizes all relevant information. In an efficient market F_{t-1} summarizes all the information concerning the expected value of S_t that is available at period t − 1. Specifically, one of the items of information available at t − 1 is the stock of information available at t − 2, and if the market is efficient, that information will be contained in F_{t-2}. If, however, F_{t-1} summarizes all available information including that contained in F_{t-2}, we should expect that adding F_{t-2} as an explanatory variable to the right-hand side of (2) will not affect the coefficient of determination and will have a coefficient that is not significantly different from zero. Equation (2″) reports the results of that regression:

$$(2'') \quad \log S_t = \underset{(.26)}{-.45} + \underset{(.08)}{1.10} \log F_{t-1} - \underset{(.08)}{.006} \log F_{t-2}$$
$$\bar{R}^2 = .98; s.e. = .46; D.W. = 1.91$$

The results in (2″) support the efficient market hypothesis.[5]

To examine further the stability of the regression coefficients during the various phases of the hyperinflation I divided the sample into two parts: "moderate" hyperinflation and "severe" hyperinflation, where the latter characterized the last nine months of the sample period.[6] A Chow test was performed on the estimates of equations (2′) and (2″) to test the equality of each and every coefficient of the two subperiod's regressions. This procedure showed that the hypothesis that the regression coefficients do not differ between the two subperiods cannot be rejected at the 95 percent level.

The results reported in this section provide support for the notion that during the hyperinflation expectations may have behaved "rationally" in the sense of Muth, and that one may use data from the foreign exchange market to measure expectations.

[3]Since the *D.W.* statistic tests for the presence of first-order autocorrelated residuals, I have also examined higher order correlations up to 12 lags; no correlation of any order was significant. To test for the hypothesis that the distribution generating the sample is normal, I have computed the studentized-range (the ratio of the range of the residuals to their standard error); the resulting statistic was 4.54—well within the interval for the 95 percent confidence level (3.57, 5.06). We thus cannot reject the hypothesis that the residuals are normally distributed.

[4]This inference, via Jensen's inequality, may be due to the choice of the numeraire. For details see J. Huston McCulloch.

[5]Following the procedure used by Eugene Fama I have also regressed $\log S_t$ on $\log F_{t-1}$ and on $\log S_{t-1}$. Since these two lagged variables are highly correlated (they actually summarize information corresponding to the same point in time), the resulting separate point estimates are inefficient; however, the sum of their coefficients did not differ significantly from unity.

[6]The moderate hyperinflation during the first 22 months of the sample period (February 1921–November 1922) was about 20 percent per month while the severe hyperinflation during the last 9 months of the sample period (December 1922–August 1923) was about 80 percent per month.

In fact, it should not be surprising that even during the turbulent period of the hyperinflation the rational expectations hypothesis might apply. It stands to reason that the larger variability of the exchange rate increases the rate of return from and the amount of resources invested in accurate forecasting.[7]

The notion that futures markets provide information on inflationary expectations need not be specific to the market for foreign exchange and in principle one could use data from other commodity futures markets. The advantage, however, of using the market for foreign exchange is that it avoids to a large extent the difficulty of distinguishing changes in relative prices from changes in the absolute level of prices – a distinction that is critical in the analysis of futures prices of specific groups of commodities. In that respect, the expectations about the future course of the exchange rate may be used as a reasonably good proxy for the expectations about the path of the aggregate price level. During the period under consideration the correlation among the various aggregate price indices and the exchange rate exceeded .99. In view of the high correlation, it seems reasonable to identify the expectations on the exchange rate with the expectations on the price level.[8]

[7]It should be noted that although on the average, the forward exchange rate is shown to be a good predictor of the future spot exchange rate and the exchange rate in turn is closely related to the price index, the forward premium on foreign exchange appears to underestimate the future rate of inflation This fact, however, need not be inconsistent with the rational expectations hypothesis. The latter requires that individuals use efficiently all available information In fact a rational expectations approach might predict that, in the absence of a previous experience with an environment of hyperinflation, while individuals learn the new structure, mistakes would occur and expectations would initially underpredict the actual course of events. For an application of the rational expectations hypothesis to the German hyperinflation see Sargent and Wallace; Sargent (1977); Evans; Aris Protopapadakis; Michael Salemi.

[8]A detailed analysis of the purchasing power parity during that period is provided in my paper (1976a).

II. The Demand for Money During the Hyperinflation

In the present section we incorporate the information on expectations into the estimation of the demand for money. We start with an analysis of the appropriate functional form.

A. *The Functional Form*

In the absence of a direct measure of expectations, estimating the demand for money involves a joint estimation of two hypotheses: (i) the specific functional form chosen for the demand for money, and (ii) the specific formulation on the process by which individuals are assumed to form expectations. Thus, for example, Cagan hypothesized that the demand for money has a semilogarithmic form and that expectations are formed adaptively as in equations (3)–(4):

$$\log\left(\frac{M}{P}\right)^d = \gamma' - \alpha\pi + u \tag{3}$$

$$\frac{d\pi}{dt} = \beta(\dot{P}/P - \pi) \tag{4}$$

where $(M/P)^d$ denotes the demand for real cash balances; M denotes the nominal money stock; P the price level; $\dot{P}/P$ and π the actual and the anticipated rates of inflation, respectively. The formulation in equation (3) reflects the assumption that during the hyperinflation, changes in the demand for money were dominated by changes in inflationary expectations so that the effects of changes in output and the real rate of interest on the demand for money may be ignored.[9] Equating the demand for real balances with the supply permits replacing

[9]This assumption seems justified in view of the small variations in output relative to variations in the anticipated inflation rate. The percentage of unemployed members of trade unions varied from 4.7 percent in February 1921 to 6.3 percent in August 1923 (see Costantino Bresciani-Turroni, p. 449) while an overall employment index varied from 95.6 percent in February 1921 to 95 percent in August 1923 (see Barro 1970, p. 1270).

$(M/P)^d$ by M/P in equation (3). Cagan's estimation procedure, which was later followed by others, amounted to generating alternative series of π corresponding to selected values of β in (4) and then choosing the value of the adaptive coefficient β which resulted in the best fit in (3).

The measure of expectations suggested in the previous section provides an independent set of data, and it makes possible an analysis of the choice of the appropriate functional form for money demand that does not depend on assumptions concerning formation of expectations. We consider two alternative functional forms: (i) the semilogarithmic form as in (3), and (ii) the double-logarithmic form as in (5).

$$(5) \qquad log \frac{M}{P} = c - \eta' \, log \, \pi + u$$

where η' denotes the elasticity of the demand for real balances with respect to the anticipated rate of inflation.

One of the difficulties of using the double-log form is that during some months, early in the sample period, the forward premium on foreign exchange was negative (reaching $-.8$ percent per month in early 1921) presumably reflecting the initial expectation that the price rise has been temporary, that the process will reverse itself and prices will return to their previous level.[10] Since the logarithm of a negative quantity is not defined, the independent variable in (5) was transformed from π to $(k + \pi)$ which henceforth is referred to as π^*. A maximum likelihood estimation of (5) along with the value of k resulted in a value of k which lies between .9 and 1.1 percent per month. For ease of exposition, in what follows we set the value of k at 1 percent per month, and thus, the coefficient η (the elasticity of the demand for real cash balances with respect to π^*) may be viewed as the interest elasticity of the demand for money.[11]

The choice of the functional form can be made by following the Box-Cox procedure of estimating the function

$$(6) \qquad log \frac{M}{P} = \beta_0 + \beta_1 \frac{\pi^{*\lambda} - 1}{\lambda} + u$$

The transformed variable $(\pi^{*\lambda} - 1)/\lambda$ has a value of $\pi^* - 1$ for $\lambda = 1$ and a value of $log \, \pi^*$ when λ approaches zero. Thus, when $\lambda = 1$, the functional form is the semilogarithmic form as in (3), while when $\lambda = 0$ the functional form is the double-logarithmic form as in (5). Since the residuals of (6) are highly correlated, the proper procedure is to perform a maximum likelihood estimation of (6) along with the values of λ and ρ (the first-order autocorrelation coefficient).[12] The likelihood ratio test indicates that the semilogarithmic form is the appropriate functional form. Since, however, the likelihood function is relatively flat, and since, as will be shown below, the autocorrelation coefficient is much lower in

[10] These regressive expectations were widespread in particular among foreign speculators who were more confident than Germans themselves in the future of Germany. These expectations account for what otherwise would have been regarded as a major paradoxical phenomenon of large foreign holdings of the rapidly depreciated Reichsmark. In fact even in October 1922 about 22 milliards paper marks were sold to foreigners through foreign banking houses; for further evidence see Bresciani-Turroni, pp. 52 and 86–87. These large foreign holdings are also consistent with the hypothesis that the forward premium is the relevant variable in the demand for money function. For a theory of expectations formation along these lines, see my paper (1975).

[11] While the anticipated difference between the domestic and the foreign rates of inflation that is measured by the forward premium π may be negative, the nominal rate of interest may not. To the extent that the relevant variable in the demand for money is "the" nominal rate of interest (which from (1) is equal to $\pi + i^*$), our estimate of k is not unreasonable in proxying "the" foreign nominal rate of interest. The value of k is also consistent with the value of the real rate of interest used in Barro (1972, p. 985). An analogous computational difficulty was faced by Barro who computed the square-root of the average rate of inflation over several time intervals and set equal to zero the few negative values which occurred in the early period (see Barro 1970, p. 1253, fn. 37).

[12] For the proper procedure as well as a description of the Box-Cox transformation, see Arnold Zellner (1971, ch. 6). For a recent application of the Box-Cox transformation for estimating the model with Bayesian techniques see Bisignano; see also Paul Zarembka; John Spitzer.

the double-logarithmic form, I will present in the following section estimates for both functional forms.

B. *Estimates of the Demand for Money and the Price Deflator*

Prior to estimating equations (3) and (5) two issues are noteworthy. The first relates to the appropriate price level to be used in deflating the nominal money stock. To a large extent the question is empirical and depends on the theory that underlies the derivation of the demand for money. The presumption, however, is that if the aggregate demand for money is dominated by households' behavior, then the relevant price index should be the consumer price index (the cost of living). The estimates of the demand using both alternative deflators, the wholesale price index, and the cost of living index, are reported below.[13]

A slightly more subtle question involves the dating of the price level. Should the money stock be deflated by the current price level or, should it be deflated by the average price that is expected to prevail during the spending period? To examine this latter question I have constructed a series of the average expected price by multiplying the current price by a factor that is equal to 1 plus the average rate of inflation that is expected to prevail over the following month; the latter, in turn, was measured by ϕ where $\phi \equiv (lnF - lnS)/2$, although a more refined measure would have related the length of the spending period to the changing length of the payments period which is known to get shorter as inflation accelerates.[14] We thus use four alternative deflators: (i) the wholesale price index, P_w; (ii) the anticipated monthly average wholesale price index, $P_w(1 + \phi)$; (iii) the cost of living index, P_c; (iv) the anticipated monthly average cost of living index, $P_c(1 + \phi)$. The various estimations reveal that during that period the distinction between the current price and the anticipated average price does not yield estimates that are significantly different from each other; I therefore do not report the detailed results for the anticipated monthly average price deflators. Table 1 reports the results of estimating the two functional forms for the two alternative price deflators.

The results in Table 1 indicate the significant difference arising from using alternative price deflators. Generally, the interest elasticity of demand for money is larger (in absolute value) when the price deflator is the wholesale price index rather than the cost of living index. The more striking fact relates to the interest elasticity of demand for money which ranges between the value of $-1/2$, as predicted by the typical Baumol-Tobin-Barro models of the transactions demand for cash, and the value of $-1/3$, as predicted by the Miller-Orr-Whalen models of precautionary and transactions demand. The semilogarithmic elasticities are somewhat lower than the estimates of Cagan, Barro (1970), and Khan (1977), and somewhat higher than the estimates of Khan (1975). As can be seen from Table 1, although the standard errors of the regressions are lower in the semilogarithmic functional form, a large fraction of the fit is attributed to the high value of the autocorrelation coefficient. In contrast, the fit of the double-logarithmic functional form is mainly attributed to the economic variables rather than to the autocorrelation coefficient.[15]

Since the estimation procedure and the approach in the present paper differ fundamentally from those of Cagan, Barro, and

[13] While Cagan, Sargent and Wallace, and Khan used the wholesale price index, Barro (1970) used the cost of living index.

[14] On the role of the endogeneity of the payments period during the German hyperinflation see Barro (1970) and Peter Garber.

[15] The maximum likelihood estimation of (6), with the cost of living index as the price deflator, results in an optimal $\lambda_1^* = .980$ with a standard error of the regression $\hat{\sigma}(\lambda_1^*) = .1091$ (compared with $\hat{\sigma}(\lambda_1 = 1) = .1093$ and $\hat{\sigma}(\lambda_1 = 0) = .1338$). When we use the wholesale price index as the price deflator the optimal λ is $\lambda_2^* = .975$ with a standard error of the regression $\hat{\sigma}(\lambda_2^*) = .1461$ (compared with $\hat{\sigma}(\lambda_2 = 1) = .1462$ and $\hat{\sigma}(\lambda_2 = 0) = .1876$). Since the transformation in (6) is applied only to the independent variable, the maximum likelihood estimate is the estimate that min-

TABLE 1—DEMAND FOR MONEY AND FORWARD PREMIUM, MONTHLY DATA: FEBRUARY 1921–AUGUST 1923

Dependent Variable	Constant	$log\ \pi^*$	π^*	$s.e.$	$\bar{R}^2$	$D.W.$	ρ
$log \frac{M}{P_c}$	6.197 (.057)	−0.358 (.028)		.134	.955	1.99	.501
$log \frac{M}{P_c}$	4.980 (.560)		−1 651 (270)	.109	.970	1.66	.967
$log \frac{M}{P_w}$	5.718 (.078)	−0.542 (.039)		.188	.959	1.97	.487
$log \frac{M}{P_w}$	4.637 (.519)		−3.316 (.362)	.146	975	1.88	.950

Note: P_c = cost of living index; P_w = wholesale price index, standard errors are in parentheses below each coefficient; ρ is the final value of the autocorrelation coefficient. An iterative Cochran-Orcutt transformation was employed to account for first-order serial correlation in the residuals. Similar results were obtained with a Hildreth-Lu transformation. The $\bar{R}^2$ is the coefficient of determination (corrected for degrees of freedom) and $s.e.$ is the standard error of the equation To verify that the serial correlation was indeed of the first-order we have examined the autocorrelation and the partial autocorrelation functions of the residuals from the above equations for 12 lags and found that the correlation among the various residuals has been removed To allow for a possible simultaneous equation bias due to the endogeneity of the forward premium the above equations were also estimated using a two-stage least squares procedure with the percentage change in the money supply and the money-bond ratio as instruments. None of the coefficients was significantly affected. Data on money and prices are from Frank Graham and Jan Tinbergen and from primary sources referred therein.

imizes the sum of squared errors. Thus, the standard error of the regression is inversely related to the likelihood function. Denoting the unconstraint parameter space by Ω and the constraint parameter space by ω, we obtain the corresponding maximum likelihood $L(\hat{\Omega})$ and $L(\hat{\omega})$. Define the ratio $\phi = L(\hat{\omega})/L(\hat{\Omega})$. The null hypothesis is rejected when ϕ is smaller than a critical value to be determined by the significance level Minus twice the difference in the logarithmic likelihood between a null and alternative hypothesis is distributed χ^2 with degrees of freedom equal to the number of restrictions imposed by the null hypothesis. Thus we have:

$$-2\ ln\ \phi = 2N[ln\ \hat{\sigma}(H_0) - ln\ \hat{\sigma}(\lambda^*)] \sim \chi^2(1)$$

where N denotes the number of observations. The likelihood ratio test indicates that at the 95 percent confidence level one cannot reject the null hypothesis that $\lambda = 1$ (i.e., that the functional form is semilogarithmic) while, at the 95 percent confidence level the null hypothesis that $\lambda = 0$ must be rejected. However, as indicated above, the superiority of the semilogarithmic form is due to the high value of the autocorrelation coefficient. For example, when we use the cost of living as the price deflator, the standard error of the regression when we include the autoregressive component of the error is .166 for the double-logarithmic form and .495 for the semilogarithmic form.

others, comparison of the results is difficult. In particular, it seems that a comparison of goodness of fit is entirely inappropriate since the other authors searched over various possible lags of the adaptive expectations so as to maximize the fit. In the present paper the expectations series were taken from an independent source of data rather than estimated jointly with the demand for money function.

We now return to discuss the issue of the appropriate price deflator. As indicated above the choice of the price deflator depends on, among other things, whether the aggregate demand for money is dominated by the behavior of households or firms. A comparison of goodness of fit of the regressions in Table 1 does not shed light on the question of the appropriate price deflator since the two sets of equations in Table 1 differ in the dependent variable. In principle, however, we may define a composite price deflator $\tilde{P}$ which depends on the wholesale price index P_w and on the cost of

TABLE 2—DEMAND FOR MONEY AND THE PRICE DEFLATOR, MONTHLY DATA: FEBRUARY 1921–AUGUST 1923
(Dependent Variable: $log\ M/P_c$)

Constant	$log\ \pi^*$	π^*	$log(P_w/P_c)$	s.e.	$\bar{R}^2$	D.W.	ρ
6.145	−0.378		0.110	.135	.956	2.01	.496
(.100)	(.042)		(.172)				
4.855		−2.144	0.298	.101	.975	1.54	.964
(.483)		(.325)	(.126)				

Note: P_c = cost of living index; P_w = wholesale price index; standard errors are in parentheses below each coefficient; ρ is the final value of the autocorrelation coefficient. An iterative Cochran-Orcutt transformation was employed to account for first-order serial correlation in the residuals; *s.e.* is the standard error of the regression. The coefficient of $log(P_w/P_c)$ yields the weight that should be given to the wholesale price index in the construction of the composite price deflator.

living index P_c according to:

$$(7) \qquad \tilde{P} = P_w^{\theta} P_c^{1-\theta}$$

Using $\tilde{P}$ as the price deflator we may estimate the parameters of the demand function along with the weights θ and $1 - \theta$. This procedure will thus provide an estimate of the relative importance of the two price indices in determining the demand for money.[16]

Consider for example the semilogarithmic form with $\tilde{P}$ as the price deflator:

$$(8) \qquad log\frac{M}{\tilde{P}} = \gamma' - \alpha\pi^*$$

Substituting equation (7) in (8) and adding an error term yields

$$(9) \qquad log\frac{M}{P_c} = \gamma' - \alpha\pi^* + \theta\, log(P_w/P_c) + u$$

which can be estimated so as to provide estimates of the weights θ and $1 - \theta$. A similar procedure may be followed for the double-logarithmic form yielding

$$(10) \qquad log\frac{M}{P_c} = c - \eta\, log\,\pi^* + \theta\, log(P_w/P_c) + u$$

Table 2 reports the results of estimating the two functional forms (9)–(10). The results indicate that in constructing the composite price index the main weight should be attached to the cost of living index. In fact, the estimate of θ (the weight of the wholesale price index) does not differ significantly from zero in the double-logarithmic form while it is about .3 in the semilogarithmic form. We may conclude, therefore, that when the choice is between these two price indices, the cost of living index seems to be the more appropriate price deflator.

C. *Stability of the Demand for Money*

The estimates in Table 1 are based on data which displayed an extraordinary degree of variation. It is important to examine whether the estimated coefficients remained stable throughout the inflationary process Since Cagan's estimates did not fit well in the last part of the hyperinflation, it was suggested that the coefficients may not have been fixed throughout the various phases of the hyperinflation (see Cagan, pp. 58–64). To examine this possibility I have split the sample into its two phases and allowed for different coefficients for the last nine months. None of the dummy variables (on the constant term and on the slope coefficient) was statistically different from zero—in fact, in most cases they were less than one standard error away from zero Thus, the hypothesis that the coefficients of the demand for money remained stable during and between both phases of the hyperinflation (the moderate and the severe

[16] For a similar procedure of estimating the weights of the various components of the wholesale price index see Garber.

TABLE 3—SHORT-RUN AND LONG-RUN DEMAND FOR MONEY, MONTHLY DATA: FEBRUARY 1921 AUGUST 1923

Dependent Variable	Constant	$\log \pi^*$	π^*	$\log\left(\frac{M}{P}\right)_{t-1}$	Long-Run Elasticity	s.e.	$\bar{R}^2$	h	ρ	Mean Lag
$\log \frac{M}{P_c}$	3.601 (.567)	−0.232 (.034)		0.418 (.091)	−0.399	.105	.973	−.539	.287	.72
$\log \frac{M}{P_c}$	3.403 (.543)		−2.003 (.231)	0.430 (.097)	−3.514	.091	.980	−1.899	.892	.75
$\log \frac{M}{P_w}$	4.707 (.640)	−0.460 (.064)		0.176 (.111)	−0.558	.184	.962	.119	.391	.21
$\log \frac{M}{P_w}$	4.648 (.593)		−3.330 (.369)	0.028 (.100)	−3.426	.149	.975	.296	.945	.03

Note: P_c = cost of living index, P_w = wholesale price index, standard errors are in parentheses below each coefficient; ρ is the final value of the autocorrelation coefficient An iterative Cochran-Orcutt transformation was employed to account for first-order serial correlation in the residuals, h is the statistic suggested by Durbin measuring the absence of first-order serial correlation in the presence of a lagged dependent variable; *s.e.* is the standard error of the regression; the mean lag is measured by $(1-\gamma)/\gamma$.

phases) cannot be rejected at the 95 percent confidence level.[17]

D. *Short-Run and Long-Run Demand Functions*

The adaptive expectations mechanism incorporates a lag of adjustment that characterizes the speed by which forecast errors are corrected. In principle one could envisage an additional lag induced by the cost of adjusting the existing stock of real cash balances to its desired level. Cagan's assumption, which was later adopted by others, was that the latter lag is insignificant. He emphasizes, however, that as a theoretical matter the two lags enter the model symmetrically and indistinguishably (see Cagan, pp. 74–77).[18]

In what follows we analyze the implications of the second of the aforementioned lags. Consider the long-run semilogarithmic demand function:

$$\log m_t^* = a + b\pi_t^* \tag{11}$$

where m^* denotes desired long-run real cash balances. Assume a partial adjustment model where the percentage rate of attaining this long-run target is proportional to the (logarithm of the) ratio of the desired level to the actual quantities:

$$\log m_t - \log m_{t-1} = \gamma(\log m_t^* - \log m_{t-1}) \tag{12}$$

It follows that

$$\log m_t = \gamma \log m_t^* + (1-\gamma)\log m_{t-1} \tag{12'}$$

where γ measures the speed of adjustment. Substituting (11) in (12′) yields

$$\log m_t = a\gamma + b\gamma\pi_t^* + (1-\gamma)\log m_{t-1} \tag{13}$$

with an analogous equation for the double-logarithmic form.

Table 3 reports the result of estimating

[17] This result holds for both functional forms of the demand for money. The fact that neither of the (mutually inconsistent) hypotheses of constant elasticity and constant semielasticity over the whole period can be rejected at the conventional confidence level reflects the "flatness" of the likelihood function. This latter phenomenon by itself is somewhat surprising given the large variation in the data during the hyperinflation.

[18] For a demonstration that the two lags lead to identical reduced form equation (if we ignore the properties of the error term), see Zvi Griliches. Roger Waud has shown that (in small samples) if both lags are present but the model assumes that only one is present, the bias in the estimates might be very significant.

the demand function with the lagged dependent variable for the two functional forms using both price deflators. Also reported in Table 3 are the values of the implied long-run elasticities and the average lag.[19] As may be seen the lagged dependent variable enters significantly only where the price deflator is the cost of living index, in which case the estimates of the speed of adjustment are .582 for the double-logarithmic form and .570 for the semilogarithmic form.[20] These estimates imply that the time it takes to complete 90 percent of the stock adjustment is about 2.7 months. The long-run semielasticity is −3.514 which, in contrast with the results in Table 1, is closer in size to the corresponding quantity in the equation using the wholesale price index. I conclude, therefore, that while the two price deflators yield very different short-run elasticities, these differences are much narrower in the long run as a result of different speeds of adjustment.

E. *Price Variability, Uncertainty and the Demand for Money*

Previous studies have suggested that the specification of the demand for money should include some measure of the degree of uncertainty as proxied by the variability of the rate of inflation (see Benjamin Klein; and Khan 1977); for a further analysis see Eden (1975, 1976). As a theoretical matter, however, the effect of the rate of inflation on the demand for money is ambiguous. On the one hand, a higher variance may raise the degree of uncertainty and thereby raise the precautionary demand. On the other hand, the variability of prices may reduce the usefulness of money as a unit of account and thereby reduce the extent to which the economy is monetized. Furthermore, the concept of variability need not coincide with that of uncertainty; surely, a path of prices which displays large variation may be associated with less uncertainty than another path of prices displaying less variation—for example, if the former is fully anticipated and the latter is not. A relevant variable, therefore, might be the variance of the prediction error.

We measure π^*—the expected rate of depreciation during period t (as perceived at period t − 1)—by the forward premium on foreign exchange, i.e., $\pi^* = \ln F_{t-1} - \ln S_{t-1}$, and we denote the actual change in the exchange rate by $\tilde{\pi} = \ln S_t - \ln S_{t-1}$ Thus, the variance of the forecast error is $\mathrm{Var}(\tilde{\pi} - \pi^*)$ which is assumed to be formed by the adaptive process. Since it is not possible to estimate the variance of a single observation, the adaptive process was applied to the first and second moments according to[21]

$$(14) \quad E_t(x) = E_{t-1}(x) + \delta(x_t - E_{t-1}(x))$$

$$(15) \quad E_t(x^2) = E_{t-1}(x^2) + \delta(x_t^2 - E_{t-1}(x^2))$$

where x denotes the forecast error. Successive substitution for $E_{t-1}(x)$ yields the known result that the expectations are weighted averages of all past observations with exponentially declining weights that sum to unity. These weights may therefore be interpreted as probabilities and hence, when the sum converges, the expectations will exist and the variance can be computed as usual:

$$\mathrm{Var}_t(x) = E_t(x^2) - [E_t(x)]^2$$

The specifications corresponding to those

[19]For an analysis of the distinction between short-run and long-run demand for money in the presence of partial adjustment see Gregory Chow. As indicated above, the present model assumes that the source of the lagged adjustment is only due to a portfolio adjustment lag and not due to lags in the formation of expectations. Edgar Feige has shown that when both lags are operative, the reduced form equation must include two lagged values of the dependent variable. To examine this possibility we have reestimated the model with a second lag of the dependent variable; in all cases the second lag did not enter significantly.

[20]These estimates of the speed of adjustment are higher than those obtained by Cagan (using the wholesale price index). Cagan did not find evidence that the speed of adjustment has risen during the later phases of the hyperinflation (see Cagan p. 60). Khan (1977) allowed the speed of adjustment to depend on the level and on the change in the rate of inflation and found that the speed of adjustment rose somewhat during the latter phases of the hyperinflation.

[21]In implementing the procedure, the initial values of $E(x)$ and $E(x^2)$ were estimated as the arithmetic averages of the first six observations.

of Table 1 were extended to include the variance as an additional independent variable and were estimated together with the corresponding adaptive expectations rule. The method of estimation was to generate different series of the variance corresponding to alternative values of the adjustment coefficient δ, and then to choose the relationship that yielded the best fit. Clearly, if the best fit is reached when $\delta = 1$, the implication is that the specified variance does not belong in the demand for money. Performing the estimation with *log* Var$(\hat{\pi} - \pi^*)$ as the added variable revealed that in all cases none of the coefficients of the variance terms was significantly different from zero, and for both specifications the best fit was reached when $\delta = 1$.[22] I conclude that the evidence from the German hyperinflation does not support the hypothesis that the variance (as computed above) belongs in the demand for money.

III. Inflationary Finance and the Money Supply Process

The analysis hitherto focused on estimates of the demand for money. To gain insights into the determinants of the inflationary process, I turn now to an analysis of the supply of money starting with a discussion of inflationary finance.

A. *Government Revenue from Inflation*

The reparation payments imposed on Germany under the Treaty of Versailles imposed a potentially heavy budgetary burden. The acceptance of the Ultimatum of London in early 1921 required a fiscal reform which was intended to generate tax receipts sufficient for the reparation payments. While these developments affected expectations, they did not by themselves induce an accelerating hyperinflation. In February 1921 only about half of the government budget was financed by floating debt, most of which was absorbed by the private sector, while the other half was financed by taxes. During that period only about 33 percent of the outstanding treasury bills were held by the Reichsbank. The budgetary burden became severe with the occupation of the Ruhr in early 1923 and the loss of important sources of tax income. The tremendous deficits induced by the massive relief payments to Ruhr workers were associated with a significant deterioration of the fiscal system and an acceleration of the hyperinflation. In August 1923 government revenue from taxes amounted to only 8 percent of the revenue generated by floating debt. The reduced attractiveness of treasury bills lowered the demand on the part of the private sector. As a result, an increasing fraction of new issues of government debt had to be monetized as it got absorbed in the portfolio of the central bank. By August 1923 more than 80 percent of the outstanding debt was held by the Reichsbank, and thus the inflationary impact of floating debt became much more significant than at the early period since it corresponded to an almost equivalent rise in the issue of notes.[23]

The gradual deterioration of the fiscal system and the increased reliance on floating debt as the prime instrument of government finance have stimulated interest in the question of what determines the amount of resources that a government can extract from an economy by means of inflationary finance (see Keynes 1923, ch. 2).[24] An ex-

[22]It should be noted that when instead of the logarithm of the variance we added the variance itself, it was never significant in the equations using the wholesale price index as a deflator. However in the semilogarithmic form of the demand for money using the cost of living as the price deflator, a maximum likelihood (while extremely flat) was reached when δ, the adjustment coefficient, reached the value of .4. In that case the coefficient of the variance was insignificant at the 95 percent confidence level, but significant at the 90 percent confidence level. It seems, however, that this result should not lead to a change of the conclusions in the text.

[23]The estimates of the shares of private and public sectors' holdings of treasury bills and of the shares of taxes and debt issue in government revenue are computed from Statistisches Reichsamt (1924, pp. 31 and 63).

[24]The semilogarithmic functional form implies that (in the absence of growth) the steady-state revenue-maximizing rate of inflation is $1/\alpha$. Previous estimates of α have suggested that the rate of money creation

amination of the path of the real value of money creation indicates that, when the wholesale price index is used to deflate the nominal addition to the money supply, real government revenue from money creation was relatively stable.[25] It is consistent therefore with the hypothesis that the rate of monetary expansion was itself endogenous to the desire to secure command over a given real amount of resources.[26]

Viewing the rate of monetary expansion in that perspective suggests that the acceleration of the rate of inflation need not indicate a situation of a self-generated inflation resulting from a violation of the familiar stability condition ($1 < \alpha\beta$) of Cagan's model (equations (3)-(4) above).[27] Under rational expectations individuals form expectations on the basis of the predictions of the underlying economic model. Therefore, expectations about the price level will be based on expectations about the money supply process and, after a learning period, the expectations about the price level will replicate (the systematic part of) that process. It follows, therefore, that in such a model expectations (or their speed of adjustment) cannot be the fundamental source leading towards instability. Rather, the speed of adjustment itself is an endogenous part of the system and instability of inflation arises only from instability of the money supply process which underlies the generation of the path of prices. Within this framework it seems natural that to understand the economics of hyperinflation one needs to examine in more detail the characteristics and determinants of the money supply process rather than concentrate mainly on the characteristics of the demand function. We turn now to discuss that question.

B. *Money and Prices: Tests of Causality*

In this section we analyze the interdependence between money and prices and

seems to have exceeded the sustained rate which maximizes government revenue. In trying to account for this Milton Friedman suggested that governments are shortsighted and behave myopically. The author (1976b) analyzed the conditions under which the present value of government revenue with monetary growth rate that exceeds $1/\alpha$ is higher than the present value of the proceeds from lower monetary growth due to the dynamic path of adjustment Evans modified Cagan's model and concluded from his new estimates that the revenue-maximizing inflation rate is much higher than was estimated in previous work and in fact during most of the period the inflation rate was below the revenue-maximizing rate Sargent (1977) suggests that the evidence itself may be an artifact due to a statistically inconsistent estimate of α. He then shows that under the general conditions which make Cagan's adaptive expectations process rational, the parameter α is not econometrically identifiable. Moreover when some restrictions are assumed on the correlation between the shocks to the demand for and the supply of money, the parameter α is usually poorly estimated. It might be noted that the approach of computing the steady-state revenue-maximizing rate of inflation leaves unanswered the question of why a reasonable government should seek to maximize its revenue by means of inflation. Moreover, the analytical framework assumes that the government has a monopoly power over the issue of notes an assumption that did not hold during the German hyperinflation. In fact the Reichsbank allowed private agencies to print and circulate their own money which was referred to as Notgeld or "emergency money." According to the president of the Reichsbank, the quantity of the various kinds of emergency money at the end of 1923 was about twice the size of the Reichsbank note circulation (see Hjalmar Schacht, p. 106).

[25]When the cost of living index is used as the price deflator, the path of real government revenue diplays a slight positive trend. The proper price deflator, however, should correspond to the patterns of government spending. A study of the budget for the period 1920-21 indicates that the proper price deflator is the wholesale price index. In that study each item of the budget, whether on a dollar basis (payments under the Treaty of Versailles) or on the wholesale price index (expenses on materials) or on the cost of living index (expenses on salaries), has been calculated in gold value The sum of these separate amounts expressed in gold was about the same as the sum obtained by converting all expenses on the basis of the wholesale price index (see Statistisches Reichsamt, 1924, p. 95).

[26]This hypothesis has been analyzed by Barro (1972), by Sargent and Wallace, and by Jacobs (1977) In principle, government revenue from money creation should be computed by using only high powered money; we ignore this distinction. It should be noted, however, that in contrast with a typical process of inflation, there is evidence that the banking system did not benefit much from the inflationary process. For references to this effect see Frank Graham, pp. 67 and 75; James W. Angell, p. 43; Bresciani-Turroni, pp 280-81 and 298. For data on the composition of the money supply see Statistisches Reichsamt, (1925, part VII).

[27]This condition is derived from a continuous-time model and should be modified when applied to em-

TABLE 4—FITTED TIME-SERIES PROCESSES-*ARIMA* (1, 1, 1)
MONTHLY DATA: FEBRUARY 1921-AUGUST 1923

Variable	Constant	AR_1	MA_1	$F(2, 28)$	$Q(12)$ P	$Q(24)$ P
$log\ M$	.027	.984	−.660	22.5	3.4	5.5
	(.060)	(.048)	(.421)		(.91)	(.99)
$log\ P_w$	.206	.522	−.898	30.1	7.6	15.8
	(.121)	(.188)	(.074)		(.48)	(.73)
$log\ P_c$	.031	.974	−.521	34.3	5.6	8.1
	(.058)	(.059)	(.176)		(.70)	(.99)

Note: All three time-series were differenced once to achieve stationarity, and were identified and estimated as integrated autoregressive moving average processes of order 1. The $Q(K)$ statistic reports the Box-Pierce test for the smallness of the whole set of the sample autocorrelations of lags 1 through K, and its approximate distribution is χ^2 with (K-p-q) degrees of freedom where p and q denote the order of the autoregressive and the moving average processes, respectively. The P values (reported below the Q-statistic) are based on the χ^2 distribution. Standard errors are in parentheses below each coefficient.

examine the existence and the direction of causality between the two series in the sense of C. W. J. Granger. We analyze the nature of the interdependence between money and prices by following two procedures. The first is an application of the Box-Jenkins time-series analysis which may be suggestive of the type of causality, and the second is an application of the more formal and rigorous procedure suggested by Christopher Sims. Starting with the Box-Jenkins approach I computed first differences of the logarithms of money and the two price indexes so as to achieve stationarity. I then estimated the sample autocorrelation functions for the three processes and identified the models as integrated autoregressive moving-average processes of order one—*ARIMA* (1,1,1).[28] The processes were then estimated and the maximum likelihood parameter estimates are reported in Table 4. The reported Q-statistics are consistent with the hypothesis that the residuals are not serially correlated and thus that the transformations reduced the observed data to white noise. In the final stage of the analysis I cross-correlated the residuals from the fitted time-series for prices (which are referred to as the prewhitened price series) with the residuals from the fitted time-series of money (the prewhitened money series).[29]

The basic idea is that these residuals are the parts of inflation and money creation that cannot be predicted from their own history. These are the "innovations" in the process and thus, if there is no significant correlation between the innovations of prices in period T and those of money in period T + K, prices and money will be said to be independent of each other in the Granger sense.

The cross-correlations are reported in Table 5, and as can be seen we must reject the hypothesis that money and prices are independent. For both price-series prices are correlated with future money one month ahead and thus suggesting that prices cause money in the Granger sense. The cross-correlation approach, however, is only valid for testing the hypothesis that prices and money are independent. Once this hypothesis is rejected, a further examination of the nature of causality requires a different procedure as the one suggested by Sims.

In implementing Sims' causality test we calculate the two-sided distributed lag regressions. To test the hypothesis that there

pirical estimates that are based on a discrete-time version of the model. See Benjamin Friedman (1975b) and Dean Dutton.

[28]It is interesting to note that for the price series for which the autoregressive coefficient is unity, the optimal forecast is the adaptive expectations rule. See John F. Muth (1960), and Sargent and Wallace.

[29]For a clear introduction to and applications of the Box-Jenkins approach, see Charles Nelson. For an application of this procedure, see Edgar Feige and Douglas Pearce.

TABLE 5—CROSS-CORRELATION OF PREWHITENED MONEY SERIES (T + K) WITH PREWHITENED PRICE SERIES (T)

Log M and	Lag. −6	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	+6	s.e.
Log P_c	−.02	−.04	−.13	−.28	.27	−.04	.49*	.43*	− 34	.03	−.12	.06	−.07	.18
Log P_w	−.21	.24	−.27	−.17	07	40*	13	.50*	−.02	−.24	.18	− 14	.011	.18

Note. Significant cross-correlations (which exceed twice the standard error) are indicated by an (*) The residuals that are cross-correlated are computed from the fitted time-series processes in Table 4

is no feedback from current inflation to future rates of money creation, we regress the current rate of inflation on past and on current and future rates of money creation. Sizable and significant coefficients on future rates of money creation lead to a rejection of the null hypothesis and permit the inference that prices cause money in the Granger sense. To test the reverse hypothesis, that there is no feedback from current rates of monetary expansion to future rates of inflation, a similar procedure is followed with a reverse distributed lag, for example, with current monetary expansion regressed on past, on current, and on future rates of inflation.

The results of these tests are reported in Tables 6 and 7 which contain two regressions for each price index: one which includes future values of the right-hand side variables and one which does not. Also reported are the F-statistics relevant for testing the hypothesis that the coefficients of future values of the right-hand side variables are zero. For example, the results in Table 6 lead to a rejection of the hypothesis that there is no feedback from current inflation to future rates of monetary expansion since the F-statistics are 13.99 and 14.44, well above the critical value of 5.72 (at the 99 percent confidence level). On the other hand, from the results in Table 7, we cannot reject the hypothesis that there is no feedback from current rates of monetary expansion to future rates of inflation measured in terms of the wholesale price index (the F-statistic is 2.55, well below the critical value of 5.72). Moreover, the absolute values of the coefficients on future rates of inflation (in terms of the wholesale price index) are much smaller than those on lagged rates of inflation. As indicated above (fn. 25), it seems that for the purpose of analyzing government revenue from in-

TABLE 6– INFLATION REGRESSED ON PAST AND FUTURE MONEY CREATION
MONTHLY DATA: FEBRUARY 1921–AUGUST 1923

Price Index	Constant	Lags −2	−1	0	+1	+2	*s.e.*	$\bar{R}^2$	*D.W.*	ρ
P_c	.025	−.342	−.416	1.087	1.049	−.350	.083	.906	1.82	−.391
	(.017)	(.246)	(.378)	(.332)	(.238)	(.064)				
	.043	.032	−.676	1.567			.120	.785	1.92	−.052
	(.031)	(.291)	(.388)	(.222)						
	F = 13.99									
P_w	.051	−.486	−.417	.335	1.866	−.405	.120	.847	2.03	−.237
	(.027)	(.353)	(.521)	(.466)	(.347)	(.093)				
	.086	.354	−1.513	1.950			.175	.645	2.05	−.119
	(.042)	(.427)	(.565)	(.322)						
	F = 14.44									

Note: P_c = cost of living index; P_w = wholesale price index; standard errors are in parentheses below each coefficient; ρ is the final value of the autocorrelation coefficient. The F-statistic corresponds to the null hypothesis that there is no feedback from current inflation to future rates of monetary creation. Critical values for $F(2, 24)$ are 3.44 (95 percent) and 5.72 (99 percent).

TABLE 7—MONEY CREATION REGRESSED ON PAST AND FUTURE INFLATION
MONTHLY DATA: FEBRUARY 1921–AUGUST 1923

Price Index	Constant	Lags −2	Lags −1	Lags 0	Lags +1	Lags +2	*s.e.*	$\bar{R}^2$	*D.W.*	ρ
P_c	−.024	−.050	.407	.339	.056	.067	.034	.975	1.87	.772
	(.037)	(.043)	(.038)	(.036)	(.031)	(.024)				
	.406	−.004	.398	.349			.045	.954	1.56	.978
	(.341)	(.054)	(.049)	(.045)						
	$F = 9.55$									
P_w	−.009	.128	.371	.231	.056	.044	.050	.948	1.53	.812
	(.062)	(.049)	(.048)	(.043)	(.040)	(.030)				
	.235	.128	.369	.219			.053	.936	1.62	.964
	(.261)	(.049)	(.043)	(.044)						
	$F = 2.55$									

Note: P_c = cost of living index; P_w = wholesale price index; standard errors are in parentheses below each coefficient; ρ is the final value of the autocorrelation coefficient. The F-statistic corresponds to the null hypothesis that there is no feedback from current rates of monetary creation to future rates of inflation. Critical values for $F(2, 24)$ are 3.44 (95 percent) and 5.72 (99 percent).

flationary finance the relevant price deflator is the wholesale price index for which the inference from Tables 6 and 7 is that during that period prices caused money while money did not cause prices in the Granger sense.[30] When inflation is measured in terms of the cost of living index, the results in Table 7 reject the hypothesis that there is no feedback from current rates of monetary expansion to future rates of inflation. Rather, they indicate a two-way causality: inflation influenced subsequent rates of money creation which in turn influenced subsequent inflation.

The analysis in this section provides insight into the determinants of the money supply process. The results concerning the interrelationship between money and prices emphasize the endogeneity of the money supply and are consistent with the hypothesis that the source of the hyperinflation, and of the accelerated rate of money expansion, was the desire (on the part of the government) to extract real resources at a rate that could not have been sustained without a continuously accelerated inflation.

[30]These results are in accord with those of Sargent and Wallace. For a further elaboration on causality tests during that period see the studies by Protopapadakis and by Salemi; for an analysis of the concept of exogeneity see Sargent (1976). The causality tests and the application of time-series analysis are discussed by Zellner (1975).

IV. Concluding Remarks

The lack of an observable variable measuring expectations has been a major difficulty in empirical work and has led to the development of various theories of the formation of expectations. In this paper I have suggested a direct measure of expectations that is based on observed data from the forward market for foreign exchange. After examining some of the efficiency properties of the market for foreign exchange, I have used this measure of expectations in estimating the demand for money and analyzing the proper price deflator during the German hyperinflation. The resultant estimates were shown to remain stable throughout the various phases of the hyperinflation. I then proceeded to analyze the money supply process and to examine the nature of the interdependence between money and prices.

Rather than summarizing the results reported in previous sections, I wish to highlight some of the issues raised by the analysis. The first concerns the role of the market for foreign exchange. Using data from that market as the basis for inference on expectations reflects the belief that even during

turbulent periods the foreign exchange market remains remarkably efficient.[31] Furthermore, the emphasis on the foreign exchange market highlights the question of what are the relevant variables measuring the anticipated cost of holding money in an open economy with flexible exchange rates. To the extent that domestic money is held as a substitute for foreign exchange, the specification of the demand for money should include the anticipated change in the exchange rate as measured by the forward premium (or discount) on foreign exchange. The evidence from the German hyperinflation illustrates the substitutability between foreign exchange and domestic money. The accelerated inflation resulted in a gradual replacement of domestic money by foreign exchange in performing the traditional roles of money as a unit of account, as a store of value, and as a medium of exchange.[32]

[31] In principle, if bonds were the relevant substitute for domestic money holdings and if the bond market were to operate efficiently during such a turbulent period, we could have used interest rates in estimating the demand for money. However, all available interest series were extremely stable up to mid-1922 and were clearly much too low to make any sense. The only series that shows some variations of interest rates is that of the day-to-day rates reported in James Angell, pp. 370-71. However, even these rates seem to be too low. When these rates were included in the estimation of the demand for money, their coefficients did not differ significantly from zero.

[32] For a description of the increasing role of foreign exchange see Graham, pp. 73-74. For the role of exchange rates in price setting behavior, see Gustav Stolper as reprinted in Fritz Ringer, p. 80. For a discussion of the role of exchange rates in wage setting behavior see Bresciani-Turroni, p. 310. Contemporary newspaper descriptions provide a vivid illustration of the increasing role of foreign exchange; for example, the *Daily Mail* of August 1923 wrote: "The mark is becoming the slave of the dollar. We have marks in our pockets but dollars in our heads." For this and other newspaper quotations see Norman Angell, ch. 8. Further evidence on the extent of replacement of domestic money by foreign exchange are reported in John Parke Young, pp. 402 and 538, according to which the value of foreign bank notes held in Germany at the end of 1923 amounted to about 1,200,000,000 gold marks while the gold value of total Reichsbank circulation amounted to only 112,100,000. The extent of the substitution resulted in numerous attempts to protect German money by legislating various controls on foreign exchange; for an account of this legislation see Statistisches Reichsamt (1924, pp. 69-70).

The second issue concerns government revenue from inflationary finance and the endogeneity of the money supply. My analysis implies that to understand the economics of inflation one needs to examine the characteristics and determinants of the money supply process rather than concentrate exclusively on the properties of the demand.

Finally, in concluding the paper, it should be noted that while this study dealt with the German hyperinflation, it is believed that the main issues raised by the analysis are also applicable to less extreme and more typical inflationary processes.

REFERENCES

James W. Angell, *The Recovery of Germany*, New York 1929.

Norman Angell, *The Story of Money*, New York 1929.

R. J. Barro, "Inflation, the Payments Period, and the Demand for Money," *J. Polit. Econ.*, Nov./Dec. 1970, *78*, 1228-63.

———, "Inflationary Finance and the Welfare Cost of Inflation," *J. Polit. Econ.*, Sept./Oct. 1972, *80*, 978-1001.

W. J. Baumol, "The Transactions Demand for Cash—An Inventory Theoretic Approach," *Quart. J. Econ.*, Nov. 1952, *66*, 545-56.

J. Bisignano, "Cagan's Real Money Demand Model with Alternative Error Structures: Bayesian Analyses for Four Countries," *Int. Econ. Rev.*, June 1975, *16*, 487-502.

G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *J. Royal Statist. Soc.*, Series B, 1964, *26*, 211-43.

Costanino Bresciani-Turroni, *The Economics of Inflation*, London 1937.

P. Cagan, "The Monetary Dynamics of Hyperinflation," in Milton Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago 1956.

G. C. Chow, "On the Long-Run and the Short-Run Demand for Money," *J. Polit. Econ.*, Apr. 1966, *74*, 111-31.

D. S. Dutton, "The Demand for Money and the Price Level," *J. Polit. Econ.*, Sept./Oct. 1971, *79*, 1161-70.

B. Eden, "Aspects of Uncertainty in Simple Monetary Models," unpublished doctoral dissertation, Univ. Chicago 1975.

———, "On the Specification of the Demand for Money: The Real Rate of Return versus the Rate of Inflation," *J. Polit. Econ.*, Dec. 1976, *84*, 1353–59.

Paul Einzig, *The Theory of Forward Exchange*, London 1937.

P. Evans, "Time-Series and Structural Analysis of the German Hyperinflation," unpublished paper, Univ. Chicago 1975.

E. F. Fama, "Short-Term Interest Rates as Predictors of Inflation," *Amer. Econ. Rev.*, June 1975, *65*, 269–82.

E. L. Feige, "Expectations and Adjustments in the Monetary Sector," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 462–73.

——— **and D. K. Pearce,** "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?," *J. Polit. Econ.*, June 1976, *84*, 499–522.

J. A. Frenkel, "Inflation and the Formation of Expectations," *J. Monet. Econ.*, Oct. 1975, *1*, 403–21.

———, (1976a) "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scand. J. Econ.*, 1976, *78*, 200–24.

———, (1976b) "Some Dynamic Aspects of the Welfare Cost of Inflationary Finance," in Ronald I. McKinnon, ed., *Money and Finance in Economic Growth and Development: Essays in Honor of E. S. Shaw*, New York 1976, 177–95.

——— **and R. M. Levich,** "Covered Interest Arbitrage: Unexploited Profits?," *J. Polit. Econ.*, Apr. 1975, *83*, 325–38.

——— **and** ———, "Transactions Costs and Interest Arbitrage: Tranquil versus Turbulent Periods," *J. Polit. Econ.*, Dec. 1977, *85*, forthcoming.

B. M. Friedman, (1975a) "Rational Expectations are Really Adaptive after All," unpublished paper, Harvard Univ. 1975.

———, (1975b) "Stability and Rationality in a Model of Hyperinflation," unpublished paper, Harvard Univ. 1975.

M. Friedman, "Government Revenue from Inflation," *J. Polit. Econ.*, July/Aug. 1971, *79*, 846–56.

P. M. Garber, "Costly Decisions and the Demand for Money," unpublished doctoral dissertation, Univ. Chicago 1976.

Frank Graham, *Exchange, Prices, and Production in Hyper-Inflation: Germany, 1920–23*, Princeton 1930.

C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, July 1969, *37*, 424–38.

Z. Griliches, "Distributed Lags: A Survey," *Econometrica*, Jan. 1967, *35*, 16–49.

R. L. Jacobs, "A Difficulty with Monetarist Models of Hyperinflation," *Econ. Inq.*, Sept. 1975, *13*, 337–60.

———, "Hyperinflation and the Supply of Money," *J. Money, Credit, Banking*, May 1977, *9*, 287–303.

John M. Keynes, *A Treatise on Money*, Vol. II, London 1930.

———, *A Tract on Monetary Reform*, 1923, Vol. IV in *The Collected Writings of John Maynard Keynes*, London 1971.

M. S. Khan, "The Monetary Dynamics of Hyperinflation: A Note," *J. Monet. Econ.*, July 1975, *1*, 355–62.

———, "The Variability of Expectations in Hyperinflations," *J. Polit. Econ.*, Aug. 1977, *85*, 817–27.

B. Klein, "The Demand for Quality Adjusted Cash Balances: Price Uncertainty in the U.S. Demand for Money Function," *J. Polit. Econ.*, Aug. 1977, *85*, 691–715.

R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

J. H. McCulloch, "Operational Aspects of the Siegel Paradox," *Quart. J. Econ.*, Feb. 1975, *84*, 170–72.

M. H. Miller and D. Orr, "A Model of the Demand for Money by Firms," *Quart. J. Econ.*, Aug. 1966, *53*, 413–35.

M. Mussa, "Adaptive and Regressive Expectations in a Rational Model of the Inflationary Process," *J. Monet. Econ.*, Oct. 1975, *1*, 423–42.

J. F. Muth, "Optimal Properties of Exponentially Weighted Forecasts," *J. Amer. Statist. Assn.*, June 1960, *55*, 299–306.

———, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

Charles R. Nelson, *Applied Time Series Analysis*, San Francisco 1973.

A. Protopapadakis, "The Role of Expectations in the Determination of Prices and Exchange Rates: Some Empirical Evidence," unpublished paper, Univ. Chicago 1976.

Fritz K. Ringer, *The German Hyperinflation of 1923*, New York 1969.

M. K. Salemi, "Hyperinflation, Exchange Depreciation and the Demand for Money in Post War I Germany," unpublished doctoral dissertation, Univ. Minn. 1976.

T. J. Sargent, "Exogeneity and Alternative Estimators of Portfolio Balance Schedules for Hyperinflation," *J. Monet. Econ.*, Nov. 1976, *2*, 511–21.

———, "The Demand for Money During Hyperinflations under Rational Expectations: I," *Int. Econ. Rev.*, Feb. 1977, *18*, 59–82.

——— **and N. Wallace,** "Rational Expectations and the Dynamics of Hyperinflation," *Int. Econ. Rev.*, June 1973, *14*, 328–50.

Hjalmar Schacht, *The Stabilization of the Mark*, New York 1927.

C. A. Sims, "Money, Income and Causality," *Amer. Econ. Rev.*, Sept. 1972, *62*, 540–52.

J. J. Spitzer, "The Demand for Money, the Liquidity Trap, and Functional Forms," *Int. Econ. Rev.*, Feb. 1976, *17*, 220–27.

Jan Tinbergen, *International Abstracts of Economic Statistics, 1919–1930*, London 1934.

J. Tobin, "The Interest Elasticity of Transactions Demand for Cash," *Rev. Econ. Statist.*, May 1966, *80*, 314–24.

R. N. Waud, "Small Sample Bias Due to Misspecification in the 'Partial Adjustment' and 'Adaptive Expectations' Models," *J. Amer. Statist. Assn.*, Dec. 1966, *61*, 1130–52.

E. L. Whalen, "A Rationalization of the Precautionary Demand for Cash," *Quart. J. Econ.*, May 1966, *80*, 314–24.

John Parke Young, *European Currency and Finance*, Vol. I, Washington 1925.

P. Zarembka, "Functional Form in the Demand for Money," *J. Amer. Statist. Assn.*, June 1968, *63*, 502–11.

Arnold A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, New York 1971.

———, "Comments on Time Series Analysis and Causal Concepts in Business Cycle Research," unpublished paper, Univ. Chicago 1975.

Statistisches Reichsamt, *Germany's Economy Currency and Finance*, Committees of Experts for the Reparation Commission, Berlin 1924.

———, *Sonderhefte zur Wirtschaft und Statistik*, "Zahlen zur Geldentwertung Deutschland 1914 bis 1923," Berlin 1925.

Budget Displacement Effects of Inflationary Finance

By JERRY GREEN AND EYTAN SHESHINSKI*

When inflation is caused by an increase in the rate of issuance of real money balances, less recourse to other sources of government revenue is necessary. This policy will therefore influence the equilibrium growth path through both the induced additional capital losses that individuals bear on their money balances and the necessary changes in fiscal policy required to balance the government budget. Though the first of these effects has received wide treatment in the literature, the second has been largely neglected.

We analyze this issue through a variety of simple monetary growth models, using alternative specifications of the government budget relation and individual savings functions. The central conclusion of all of these models is the tendency for inflation to increase capital intensity in the absence of any effects on portfolio proportions or the savings rate induced by changes in interest rates. We also present some numerical results for these models using parameters related to the current *U.S.* situation. Typically, a 1 percent increase in the permanent rate of inflation will produce a capital stock that is 2–4 percent larger.

The following symbols will be used:

Tax Parameters:

- τ_1 = corporate tax rate on real profits
- τ_2 = corporate tax rate on inflation induced profits
- τ = corporate tax rate (used when $\tau_1 = \tau_2$ is assumed)
- ρ = tax rate on labor income
- T = total real taxes collected, per capita

Rates and Proportions

- n = population growth rate
- γ = government budget as proportion of real output per capita
- σ = savings rate out of disposable income
- Λ = fraction of wealth held in money balances

Commodities

- Y, F = total national income (F is used as function of capital and population)
- K = total real capital stock
- N = total population
- y, f = per capita national income (f is used as function of capital per capita)
- k = per capita real capital stock
- m = per capita real money balances
- l = per capita real government bonds
- e = per capita government real expenditures
- d = per capita disposable income

Prices

- i = nominal interest rate
- π = rate of inflation
- r = real interest rate, real before tax return to capital
- w = real wage
- P = price of output
- W = nominal wage

I. Introduction

A. *Models of Taxation and Inflation*

The literature of monetary growth theory has been consistently concerned with the ef-

*Harvard University, supported by National Science Foundation Grant SOC71-03803; and Hebrew University of Jerusalem, supported by National Science Foundation Grant SOC74-11446 at the Institute for Mathematical Studies in the Social Sciences, Stanford University. We are grateful to Robert King and the managing editor of this *Review* for helpful comments and important corrections.

fects of inflation on long-run economic equilibrium. Nonmonetary one-sector models[1] have been extended to treat the case in which governments use lump sum taxation and inflationary monetary policy to finance their expenditures.[2] In these models, inflation influences capital accumulation by changing the desired composition of portfolios as the rates of return to holding monetary and nonmonetary assets diverge. Recently, Martin Feldstein has generalized these models to include other types of taxation and a savings behavior derivable from the life cycle hypothesis.

He sets up a full-employment model of the standard type in which there is a single good, and money is the only nonphysical asset. Competitive behavior is assumed throughout. Labor grows at a constant rate n, and is supplied inelastically. Production is assumed to be in accordance with a neoclassical constant returns to scale technology. The corporate tax enters the profit-maximization calculus as follows.

A firm that is employing N units of labor at wage rate W, and K units of capital produces a flow $Y = F(N,K)$ of output, measured in the same units as capital. All capital is financed by debt denominated in monetary units, which pays a nominal interest rate i. Let P be the price of output. Therefore, at any instant of time, the firm's nominal cash flow is equal to $PY - WN - iPK$. In addition, the stock of capital, which is assumed not to depreciate, is increasing in nominal value at the rate of price increase π. These two components of profit can be taxed at different rates: τ_1 for cash flow and $\tau_1 - \tau_2$ for the inflationary inventory revaluation.[3] Therefore, after-tax profit is

$$(1) \quad (PF(N,K) - WN - iPK)(1 - \tau_1) + (\pi PK)(1 - \tau_1 + \tau_2)$$

Defining the real wage as $w = W/P$, we have that maximal profit is attained when

$$(2) \quad F_N = w$$

$$(3) \quad (F_K - i)(1 - \tau_1) + \pi(1 - \tau_1 + \tau_2) = 0$$

Using the assumption of constant returns to scale, and letting $f = F/N$ and $k = K/N$, we have

$$(4) \quad f - kf' = w$$

$$(5) \quad f' = i - \frac{1 - \tau_1 + \tau_2}{1 - \tau_1}\pi$$

Feldstein examines the dependence of the steady state on the two components of the tax rate. To simplify the analysis and help us focus on some other issues, we will assume that the nominal capital gains are untaxed. Thus $\tau_1 = \tau_2$, and (dropping the subscript on τ)

$$(6) \quad f' = i - \frac{1}{1 - \tau}\pi$$

Corporate after-tax profits in the competitive equilibrium are therefore

$$(7) \quad (f - w - ik)(1 - \tau) + \pi k$$

which is equal to zero, using (4) and (5), as one would expect by the constant returns postulate. However, it is interesting to note that before tax, the profits

$$(8) \quad f - w - ik + \pi k$$

are negative. They are, in fact,

$$(9) \quad \frac{-\tau\pi k}{1 - \tau}$$

which is the *subsidy* received by firms because the "taxable" component of "profit," $f - w - ik$, is actually a loss.

Personal income in the Feldstein model exceeds total product by exactly the amount of these losses. It is $w + rk$ on a per capita basis. Individuals are assumed to hold money according to the relation

$$(10) \quad m = \Lambda k$$

where m is the real money stock per capita. Feldstein assumed that Λ varies with the real interest rate.

The government is assumed to spend a

[1]See Robert Solow.

[2]See James Tobin, David Levhari and Don Patinkin, Duncan Foley and Miguel Sidrauski, and Jerome Stein.

[3]Feldstein writes net profit as $(1 - \tau_1)(y - wl) - (1 - \tau_1)rk - (1 - \tau_1)\pi k + \tau_2\pi k$ inventory profit. This is equivalent to our formulation.

ixed proportion γ of real output. The total government budget is therefore

$$\gamma f + \frac{\tau \pi k}{1 - \tau}$$

since firms must be subsidized from general revenues. The sources of these revenues are a personal interest income tax,[4] a lump sum tax, and inflationary finance. Thus the government's budget equation is[5]

$$\gamma f = T - \frac{\tau \pi k}{1 - \tau} + (\pi + n)m \tag{11}$$

since inflation is caused by the issuance of real money balances in excess of the rate of population growth in the steady state.

Real disposable income is personal income net of tax minus the losses on money holdings due to inflation, πm. This is

$$(1 - \gamma)f + mn \tag{12}$$

It is assumed that a fraction, σ, of this is saved and invested according to (10) in the two assets. Feldstein allowed σ to depend in a general way on the real net interest rate.

The steady-state equation is

$$\sigma((1 - \gamma)f + n\Lambda k) = n(1 + \Lambda)k \tag{13}$$

It is important to note that τ and π affect the steady-state value of k only through the values of σ and Λ which depend on the net rates of return. Therefore, as Feldstein derives, the effects of changes in inflation on the steady-state capital intensity and real money balances occur only because savings and liquidity preference are interest responsive to these parameters. His analysis suggests that the effects of savings are likely to dominate those of liquidity preference since the latter operate only through money holdings which empirically form a small proportion of total wealth.

With $\sigma = .1$, $\pi = .05$, $\Lambda = 1/40$, $\gamma = .24$, $n = .01$, and a Cobb-Douglas technology with capital's share at .25, the equilibrium is characterized by a real rate of interest of 3.28 percent. At $\tau = .5$ total personal tax collections are 60 percent of total output, although real government expenditures are only 24 percent. The difference consists of 38 percent of total output, representing subsidies to firms minus 20 percent of output which is inflationary finance.

In the absence of savings or portfolio sensitivity to interest rates, k will be a constant. The change in lump sum taxation with respect to the inflation rate can be seen from (11) to be

$$\frac{dT}{d\pi} = \frac{\tau k}{1 - \tau} - m \tag{14}$$

At a 50 percent corporate tax rate which corresponds to the current *U.S.* situation, and with $\Lambda \approx 1/40$, this is clearly positive. Thus rather than acting as a substitute for other sources of government revenues, inflationary finance creates the need for higher personal taxes!

However, higher inflation rates increase pretax personal income because of the higher level of i given by (6). But it is nevertheless true that tax collections take an increasing share of personal income net of inflationary capital losses. At $\tau = .5$ we have equation (15). Since $\Lambda \approx 1/40$, $n \approx .01$, and $\gamma \ll .5$, the derivative in (15) is clearly positive.

Because these effects of inflation on taxation are somewhat perverse, we are led to study models in which the revenue-generating aspect of inflationary finance can be separated from its other impacts. In these models it will be seen that inflation is not neutral even when savings and portfolio effects are neglected. We introduce the possibility of government borrowing and taxation on all forms of income at constant marginal rates, instead of lump sum personal taxation and interest income taxes only.

Section IB considers a model directly analogous to Feldstein's in which the savings base is broadened to include real government expenditures net of subsidies to firms, and in which total government spending rather than net expenditures only

[4] θ_1 and θ_2 in Feldstein.

[5] Here T is total personal taxes. Feldstein's T is net taxes inclusive of subsidies to firms. But he wrote personal income directly as $f - T$, which is $w + rk - T$ in our notation.

$$(15)\quad \frac{d\left(\frac{T}{w + ik - \pi m}\right)}{d\pi} = \frac{f(1 - 2\gamma - \Lambda(1 - \gamma)) + \pi k(1 - \Lambda) + (2 - \Lambda)(\pi - \Lambda(\pi + n))}{(w + ik - \pi m)^2}$$

are held at a fixed fraction of total output. We show that this model has all of the qualitative properties of Feldstein's model.

Section II treats the two possibilities that the savings base can be broadened without changing the government expenditure rule, and vice versa. These models still have the feature that lump sum taxation is used to balance the government budget. They are not exactly symmetric in their properties. These differences are explored, and the magnitudes of the inflationary effects are studied using numerical examples with reasonable parameter values.

In Section III we treat a variety of models in which government borrowing replaces lump sum taxation. The results of these are compared, both analytically and numerically. Under reasonable hypotheses about the parameters of the system, inflation will tend to increase capital intensity and decrease the steady-state value of government debt per capita in the absence of life cycle savings variation or changes in liquidity preference.

In Section IV we study a model in which firms hold money. Section V covers a model with neither borrowing nor lump sum taxation, in which the rate of inflation is endogenously determined by the tax and spending parameters.

B. *A Broader Savings Base and Proportionality of Total Government Spending to Output*

If real government purchases are used to provide private goods on a public basis, we might expect that savings out of net disposable income will respond positively to this activity. The simplest hypothesis is that real government purchases are perfectly substitutable for net disposable income in the savings base, and that the fraction of their sum saved is a constant.

Total government spending is divided into real purchases e, and subsidies to firms $\tau\pi k/(1 - \tau)$. The sum of these is a fixed fraction of real output. The government budget equation is therefore

$$(16)\quad \gamma y = e + \frac{\tau}{1 - \tau}\pi k = T + (\pi + n)m$$

The savings base can be written as

$$(17)\quad d + e = w + rk - \pi m - T$$

$$= w + rk - \frac{\tau}{1 - \tau}\pi k + nm \quad \text{(using (16))}$$

$$= f + n\Lambda k \quad \text{(using (10), (4))}$$

Therefore the condition for a steady state is

$$(18)\quad \sigma(f + n\Lambda k) = n(1 + \Lambda)k$$

Hence, when σ and Λ are constant, as we shall be supposing throughout, k is independent of both the rate of inflation and corporate taxation.

Equation (18) is not exactly the same as (13). However, due to the broader savings base we might expect σ to be lower in this case. Since savings are about 10 percent of disposable income, the comparable figure is 7.8 percent of disposable income plus government purchases. Using the other parameters as given in Section IA above, the equilibrium level of the capital stock is consistent with a real rate of interest of 3.28 percent. At a zero-corporate tax rate, the share of the savings base attributable to government spending is 24 percent. As τ rises, firms' losses increase and therefore the share of total government revenues going into real purchases falls. Since the savings base is constant this change induces an

increase in real disposable income. Personal interest income is rising since r increases with τ and all other components of disposable income depend only on k, which is a constant.

II. Budget Displacement Effects with Lump Sum Taxation

A. *Government Purchases Proportional to Output-Broad Savings Base*

In the models of the last section it was shown that inflation could not affect capital intensity, even in the presence of corporate taxation, unless savings or liquidity preference depend on the rate of return. This is due to the fact that the savings base is independent of the rate of inflation and that steady-state capital intensity is determined completely by the equality of savings and investment. In Feldstein's model, savings depend on real output plus the rate of increase in the real money stock minus real government expenditures. In our model with a broader base, the last term is omitted and one would expect a correspondingly lower average savings rate. The independence of the savings base in both models is due to the compensating changes in lump sum taxation.

In the first case, an increase in the rate of inflation increases losses made by firms, which are subsidized through higher lump sum taxes. However, perceived personal interest income increases by exactly the increase in taxation plus inflationary losses on real money balances, since the real rate of interest on the fixed capital stock increases by more than the rate of inflation. In the second case the gain in personal interest income is offset instead by the fall in real government expenditures, which are assumed to be a perfect substitute in the savings base. The level of taxation necessary to maintain a balanced budget is a constant. One can then see that by using a definition of the savings base that is compatible with the government's expenditure rule, the independence property would be maintained.

Therefore, one is led to consider other cases. For example, we could include government expenditure in the savings base, and assume that they are a constant fraction of real output. That is, we use Feldstein's specification of the government budget equation, but enlarge the savings base as in the second model of Section I.

Here, disposable income is defined by

$$d = w + rk - \pi m - T \tag{19}$$

and the government's budget equation is

$$\gamma y = e = T + (\pi + n)m - \frac{\tau}{1-\tau}\pi k \tag{20}$$

With a savings base of $d + e$, the steady-state equation becomes

$$\sigma\left(w + rk + mn - \frac{\tau}{1-\tau}\pi k\right) = n(m + k) \tag{21}$$

Using the relations for the firm's equilibrium (4) and (6) and the liquidity preference relation (10), this becomes

$$\sigma(f(k) + n\Lambda k) = nk(1 + \Lambda) \tag{22}$$

which is identical to the equilibrium condition in the second model of Section I.

This reflects the fact that real government expenditures are equal to taxes plus inflationary finance minus subsidies to firms in both cases. The broader savings base with complete substitutability between government and personal disposable income removes the influence of the government expenditures on savings.

B. *Government Budget Proportional to Output-Narrow Savings Base*

However, the situation would be markedly different if we were to take the narrower savings base of Feldstein's paper together with a government budget equation in which total expenditures, including subsidies to firms, are proportional to real output.

The government's budget equation is

(23) $\gamma y = T + (\pi + n)m$

which, when combined with (14), (16), and (10), gives the steady-state relation

$$(24)\quad \sigma(1-\gamma)f + \frac{\tau}{1-\tau}\pi k + n\Lambda k = n(1+\Lambda)k$$

Therefore, as long as the corporate tax rate is different from zero, the rate of inflation will affect the real variables in the steady state. A higher rate of inflation will decrease the level of lump sum taxes necessary to balance the budget at the same level, the magnitude of this effect being exactly equal to the inflationary finance created. Disposable income is therefore increased by exactly the increase in personal interest income as can be seen from the above remarks and equation (19). Savings increase because of this, although real government purchases are lower. The new equilibrium will therefore occur at a higher level of capital per head, and a higher real output. On the other hand, since the share of output going to government expenditures falls, no firm welfare conclusions can be drawn without a specification of individuals' tastes for alternative forms of income.

This can be seen from differentiating (24) to obtain

$$(25)\quad \frac{dk}{d\pi} = \frac{-\sigma\tau k}{1-\tau} \cdot \left(\frac{1}{\sigma(1-\gamma)f' + \frac{\sigma\tau\pi}{1-\tau} + \sigma n\Lambda - n(1+\Lambda)}\right)$$

Combined with the equilibrium condition, this is

$$(26)\quad \frac{dk}{d\pi} = \frac{-\tau k}{1-\tau}\left(\frac{1}{(1-\gamma)\left(f' - \frac{f}{k}\right)}\right)$$

which is positive by the concavity of f.

Differentiating (19) and using (26), it is easy to see that the steady-state aggregate consumption level will be increased. Differentiating (20) it can be shown that

$$(27)\quad \frac{de}{d\pi} = \frac{-\tau k}{(1-\gamma)(1-\tau)}\left(1 + \frac{e/k}{\left(f' - \frac{f}{k}\right)}\right)$$

which can have either sign. Initial forces tend to decrease e but if the elasticity of output with respect to input is sufficiently small, the resulting increase in k may offset the budget displacement effect. Even in this case, however, welfare is not necessarily improved in the new steady state because the intergenerational distribution of output is altered due to changes in the real rate of interest.

Using a Cobb-Douglas production function with capital's share set at .25 and parameter values of $\sigma = 1$, $\gamma = .24$, $\pi = .05$, $\tau = .5$, $\Lambda = .025$ (which roughly correspond to the current *U.S.* experience), the change in the capital-labor ratio induced by a 1 percent increase in the inflation rate can be computed from (26) to be 15.4 percent of its previous equilibrium value. The change in the real rate of interest is given by

$$(28)\quad \frac{dr}{d\pi} = f''\frac{dk}{d\pi} + \frac{\tau}{1-\tau}$$

which is 0.8 percent per 1 percent increase in π. This should be compared with the comparable expression obtained by Feldstein, which yields a 1 percent increase (in the absence of savings and portfolio effects) for the same parameter values. The budget displacement effect therefore mitigates the induced increase in the real interest rate found by Feldstein. If, however, government follows a policy through which real purchases are not decreased to the full extent of the additional subsidy required for firms, this effect will be correspondingly smaller. In any case it is likely to be much larger than a pure liquidity preference effect (see Tobin), which is probably less than 0.01 percent.

In the models of this section changes in the rate of inflation will not effect any real variables in the absence of corporate taxation. This is because disposable income varies only due to changes in the real rate of

interest, which remains unaffected without corporate taxation. Lump sum taxation offsets the losses on real money balances due to the increased inflation. It is therefore of interest to study such models in cases where the government issues debt, rather than changes taxes, to balance the budget when the rate of inflation is altered. The private sector then bears the burden of inflationary finance immediately. Only over time will these forces cause a compensating change in disposable income through the effect of debt service costs in the government budget. Introducing an additional asset in this way allows us to study models in which monetary policy can be analyzed without making other compensating changes in the government's actions. This is the topic of the next section.

III. Budget Displacement Effects with Public Debt

A. *Government Purchases Proportional to Real Output*

In this section we study models in which public debt is a perfect substitute for the obligations of firms in individual's portfolios. We neglect corporate taxation for simplicity. We will demonstrate that the real economic variables of the system are affected by inflation, even in the absence of a corporate tax—which was not the case in the previous models studied.

We assume that the government taxes wage income and interest income from both corporate and public debt at the same rate, ρ. Let l represent the real value of the stock government debt per capita, and $\dot{L}$ be the rate of issuance of new government debt at any instant of time, denominated in units of money.

In the presence of inflation the real value of government debt held by any individual is falling at any instant of time. We assume in this section that the tax laws allow a full deduction of these losses.[6] Since individuals regard corporate and public debt as perfect substitutes, the rate of return on these assets is equal. Disposable income is therefore

$$d = [w + r(k + l)](1 - \rho) - \pi m \tag{29}$$

Maintaining the spirit of the portfolio condition of previous sections, we assume

$$m = \Lambda \cdot (l + k) \tag{30}$$

where Λ is a constant.

If we suppose that the government purchases a constant share of real output, the government's budget equation is

$$\gamma y = \rho(w + r(k + l) - (r + \pi) + \frac{\dot{L}}{PN} + \frac{\dot{M}}{PN} \tag{31}$$

The first term on the right-hand side is real tax collections, noting that a loss-offset on both types of debt is allowed. The second term is the real debt-service paid by the government. The rate of interest is the real rate plus the rate of inflation. The fall in the value of government debt allows further borrowing at every instant, to keep the real stock of debt per capita constant. The final two terms are the real values of currently issued bonds and outside money.

In the steady state

$$\frac{\dot{L}}{L} = n + \pi \qquad \frac{\dot{M}}{M} = n + \pi \tag{32}$$

so that the real levels of debt and money are constant in per capita terms. Substituting these relations in (31) we have

$$\gamma y = \rho(w + r(k + l)) - (r + \pi)l + (\pi + n)l + (\pi + n)\Lambda(l + k) \tag{33}$$

or

$$\gamma y = \rho(w + rk) - ((1 - \rho)r - n)l + (\pi + n)\Lambda(l + k) \tag{34}$$

In the absence of corporate taxation or subsidization, firms will borrow up to the

[6] We have also recomputed our results under the assumption that these losses can be deducted from the tax base. This lowers the steady-state capital-labor ratio, but has no significant effects on the comparative statics of the system.

point where

$$f'(k) = r \tag{35}$$

and will hire labor as before, so that

$$f(k) - kf'(k) = w \tag{36}$$

Therefore, the government's budget equation becomes

$$\gamma y = \rho y - ((1 - \rho)r - n)l + (\pi + n)\Lambda(l + k) \tag{37}$$

Together with the steady-state condition

$$n(1 + \Lambda)(l + k) = \sigma\{[w + r(k + l)](1 - \rho) - \pi m\} \tag{38}$$

This determines the behavior of the steady states of this system as the rate of inflation is varied.

Unfortunately this system is likely to be unstable for typical values of the parameters, although a complete analysis would require a specification of nonsteady-state behavior in the commodity and asset markets in the presence of inflation, which is beyond the scope of this paper. The potential for instability can be seen as follows: The initial increase in π causes the issuance of new bonds $\dot{L}$ to fall, as can be seen from (31), since at that instant both budget and debt service levels are fixed by the historically given stocks. The magnitude of this decrease is 1 percent of the nominal money stock for each percent of additional inflation. Nominal disposable income goes down by 1 percent of the nominal money stock, which can be verified by multiplying (29) by the price level. Savings decrease by less than this, since $\sigma < 1$. Since the government supplies bonds inelastically at the market rate of interest, the fall in supply is greater than that in total savings so that the quantity of real output channeled into capital formation initially increases with the higher inflation rate. On the other hand, the budget equation (37) can be rewritten as

$$0 = (\rho - \gamma)f(k) - ((1 - \rho)r - n - (\pi + n)\Lambda)l + (\pi + n)\Lambda k \tag{39}$$

from which we see that if the coefficient of l is positive, a *higher* value of debt will have to be maintained at each level of capital intensity to balance the budget when inflation increases. Moreover, since the system begins to accumulate capital at a higher rate when population is growing, instability will surely result whenever $\rho > \gamma$ as well. This condition is equivalent to the fact that debt service net of taxes is greater than the level of deficit finance. The current *U.S.* data do not give direct evidence on this matter but it must be remembered that the deficit should be calculated on a full-employment basis in a steady-state situation. Taking this into account, the indicated inequality is almost surely valid. The coefficient of l, $(1 - \rho)r - n - (\pi + n)\Lambda$, is positive in any steady state with relatively moderate inflation and a savings propensity 10 or 15 times the rate of population growth. For these reasons, due to the instability of the system we believe that the comparative statics of this case are likely to be misleading. We will therefore concentrate on an alternative specification of the government budget equation related to that studied previously.

B. *Budget Proportional to Real Output-Broad Savings Base*

We will assume that instead of controlling purchases of real output, the government policy is to keep total spending, including net debt service, at a constant proportion of national product. Specifically, letting e represent government purchases of output per capita we have

$$\begin{aligned}\gamma y &= e + (1 - \rho)rl \\ &= \rho f + nl + (\pi + n)\Lambda(l + k)\end{aligned} \tag{40}$$

We will assume further that savings are proportional to disposable income plus government purchases of real output. Thus the steady-state condition becomes (using (29), (36), (38), and (40))

$$\begin{aligned}n(l + k)(1 + \Lambda) &= \sigma(d + e) \\ &= \sigma(f + nl + n(l + k)\Lambda)\end{aligned} \tag{41}$$

This savings assumption is compatible with

$$\frac{dk}{d\pi} = \frac{\Lambda(l + k)(n - f/k)}{n(f' - f/k)(l - \rho + \gamma)} \tag{42}$$

$$\frac{dl}{d\pi} = \frac{\Lambda(l + k)^2(\sigma(f' + nL) - n(1 + L))}{n\sigma(f' - f/k)k(1 - \rho + \gamma)} \tag{43}$$

the inclusion of net debt service in the government budget relation. The case of a narrower savings base will be treated later in this section.

Differentiating totally with respect to π, l, and k, and substituting the solutions of the equilibrium equations, we obtain (42) and (43). The denominators are negative by the concavity of the production function. The numerator of (42) is negative and that of (43) will be negative provided that

$$n > \sigma f' \tag{44}$$

which is assured in the steady state of a Solow-type one-sector model and is valid for our typical parameter values as well.

For example, using the parameters of Section II,[7] taking a Cobb-Douglas production function with capital's share equal to .25, $\sigma = .078$, $\gamma = .24$, $\pi = .05$, $\Lambda = .025$, and assuring further that $\rho = .22$ (instead of the lump sum taxation of Section II), we can calculate that the steady state is characterized by an interest rate of 3.6 percent.

When the inflation rate increases by 1 percent, the equilibrium level of the real capital stock increases by 3.44 percent (using (42)) and the equilibrium level of real bond holdings decreases by 0.4 percent (using (43)). This change in the capital stock reduces the equilibrium interest rate by 0.01 percent.

These are in contrast to the model of the end of Section I which was identical except for the presence of lump sum taxes instead of bond sales in the government budget equation. There we found no influence of inflation on the steady-state real variables, and an equilibrium real interest rate of 3.28 percent. Thus the real capital stock in a model with lump sum taxation under the budgetary and savings assumption we are using is equivalent to that in a model with debt finance under a much higher inflation rate.

[7]We have assumed $\sigma = .078$ to maintain comparability with other sections. We have also assumed

$$\frac{d}{d+e} = \frac{1.508}{1.918} \approx .78$$

so that

$$\sigma_{wide} = \frac{d \cdot \sigma_{narrow}}{d+e} \approx .078$$

C. *Budget Proportional to Real Output–Narrow Savings Base*

Section IIB discussed a model comparable to that of Section IA, in the sense that the savings base and the government's budget equation were the same, but government debt replaced lump sum taxes as the residual variable in the budget equation. In this part we analyze a model with the same budget specifications, but disposable income is now the only component of the savings base. Disposable income is defined as in (29) so that with total savings equal to σ times this, the steady-state equation is

$$n(1 + \Lambda)(l + k) = \sigma(1 - \rho)f + (1 - \rho)f'l - \pi\Lambda(l + k) \tag{45}$$

We repeat the government's budget equation for convenience:

$$0 = (\rho - \gamma)f + nl + (\pi + n)\Lambda(l + k)$$

Differentiating this system totally and substituting the equilibrium expressions obtained by eliminating l from the government budget equation and using (45), we obtain (46) and (47).

$$\frac{dk}{d\pi} = [\Lambda(l + k)^2 f'] \div [(f' - f/k)k(\sigma n - f'(\rho - \gamma)) + f''kl(n - f(\rho - \gamma)/k)] \tag{46}$$

$$(47)\quad \frac{dl}{d\pi} = -\Lambda(l+k)^2\left[-\frac{n}{1-\rho} + \left(\sigma + \frac{\rho-\gamma}{1-\rho}\right)f' + f''l\right] + [(f' - f/k)k(\sigma n - f'(\rho-\gamma)) + f''kl(n - f(\rho-\gamma)/k)]$$

The denominator can be shown to be negative whenever

$$(48)\quad \gamma > \rho$$

Even if this were violated, the negativity would be preserved unless σ were unrealistically low. Thus $dk/d\pi < 0$ under our assumptions.

The numerator of (47) will be positive under the same conditions provided

$$(49)\quad n > \sigma f'$$

which is surely valid for economies in which we are interested. Therefore $dl/d\pi < 0$.

Using, for illustrative purposes, the same parameters as those taken in Section IIIB, we find that the equilibrium capital-labor ratio produces a real rate of interest of 3.6 percent and the equilibrium ratio of real capital to holdings of government bonds is 8.2:1. The value of $dk/d\pi$ is -6.9 percent per 1 percent increase in the inflation rate. Real bond holdings decrease by 10.2 percent of their equilibrium value per 1 percent increase in inflation. This means that the real interest rate will respond to a 1 percent increase in inflation by rising .2 percent.

Comparing this to the model at the end of Section II, which was identical except for the presence of lump sum taxation instead of government borrowing, we see that the equilibrium interest rate is much higher here due to the fact that some wealth is channeled away from real capital formation. The former model had an interest rate of 1.7 percent at a corporate tax rate of 50 percent. At a zero-corporate tax rate the former model would have had an interest rate of 3.3 percent. Here the equilibrium interest rate is 3.6 percent.

These figures can be explained as follows: since firms make losses in the presence of corporate taxation, a lower tax rate induces lower subsidies at each fixed rate of inflation. The loss of these subsidies forces some firms out of business reducing the capital stock. Moreover, in the presence of debt finance, part of wealth (about 1/9 with these parameters) is held in this form, further reducing the equilibrium stock of productive capital.

IV. A Model with Firms Holding Money

One of the most striking disparities between the monetary growth models presented above and the real world is the fact that most money is in reality held by firms. We will show that the basic results of these models are preserved in a model analogous to that of Section IIIc, with such a modification. The simplest assumption paralleling that used above is that firms must keep real money balances proportional to capital according to

$$(50)\quad m = \Lambda k$$

and individuals hold government debt, on which a full offset for inflation produced capital losses is allowed. We will use the narrow savings base—but the results would be essentially unchanged with the wide base and corresponding modifications in the levels of the assumed parameters. For simplicity, and to isolate the effects of the change in the ownership of money balances on the system, we will assume no corporate taxation.

The rate of inflation affects firms' choices of capital intensity through the fact that capital losses on real money balances are causing them to economize on money and hence on capital, which is a complementary input. In fact, one can regard these losses as a type of depreciation since the technological justification for (50) is presumably a production function of the form $y = f(\min(m\Lambda, k))$.

Firms' profits are given by

$$(51)\quad y - w - ((1+\Lambda)i - \pi)k$$

in per capita terms. The first-order condition for a maximum is

$$(52)\quad f' = r(1+\Lambda) + \pi\Lambda$$

We assume that the government budget is proportional to $(y - \pi\Lambda k)$, which yields

$$(53)\quad \gamma(y - \pi\Lambda k) = \rho(f - \pi\Lambda k) + nl + (\pi + n)\Lambda k$$

The justification for this is that the net output of the economy is really the gross output y minus the real savings channeled into money that would be necessary to maintain a constant output level—recall the technological specification made above. This parallels the budget justification of Sections IB, IIB, and IIIB and C. Disposable income includes wages plus interest on government and corporate debt. The latter is proportional to $m + k$ since firms must finance their money holdings as well as their capital through borrowing. Thus using (51),

$$(54)\quad d = (1 - \rho) \cdot \left(f - \pi\Lambda k - \left(\frac{f' - \pi\Lambda}{1 + \Lambda}\right) l\right)$$

Finally, the steady-state equation is

$$(55)\qquad \sigma d = n(k + m + l)$$

Substituting (54) into (55), then differentiating the result and (52) totally with respect to π, l, and k at the equilibrium values, we find (56) and (57).

$$(56)\quad \frac{dk}{d\pi} = \left[(1 + \gamma - \rho)\Lambda kn + (1 + \gamma - \rho)\Lambda k\sigma(1 - \rho)r - n\sigma(1 - \rho)\Lambda\left(k - \frac{l}{1 + \Lambda}\right)\right] \div \left[\frac{f}{k} - f'\right](\gamma - \rho)(n - \sigma(1 - \rho)r) - n\sigma(1 - \rho) + \frac{nl\sigma(1 - \rho)f''}{1 + \Lambda}$$

$$(57)\quad \frac{dl}{d\pi} = \left[(1 + \gamma - \rho)\Lambda k(1 + \Lambda)(-n + \sigma r(1 - \rho)) + (1 - \rho)\sigma\Lambda(-k(1 + \Lambda) + l)\left((\gamma - \rho)r - \frac{(\pi + n)\Lambda}{1 + \Lambda}\right) - \sigma(1 - \rho)\frac{f''l}{1 + \Lambda}(1 + \gamma - \rho)k\Lambda\right] \div \left[\frac{f}{k} - f'\right](\gamma - \rho)(n - \sigma(1 - \rho)r) - n\sigma(1 - \rho) + \frac{nl\sigma(1 - \rho)f''}{1 + \Lambda}$$

One can show, using the equilibrium conditions, that sufficient conditions for the denominator to be negative are

$$(58)\qquad \gamma - \rho > 0$$

$$(59)\qquad \sigma > \frac{\gamma - \rho}{1 - \rho}$$

which are analogous to the conditions of Section III.

Numerically, using the same parameter values as in Section III, we find that the original steady-state real interest rate is 3.65 percent. The induced change in capital per 1 percent change in the inflation rate is -0.19 percent of the original capital stock.

V. Endogenous Inflation: Effects of Tax and Budget Changes

Suppose that the government finances its budget by nonlump sum taxes and without any borrowing. If real government purchases are a fixed proportion of real output, then the rate of inflation is endogenously determined in the system so as to satisfy the government's budgetary needs. This can be seen as follows.

We assume for simplicity that there are no corporate taxes, and that a full loss offset on capital losses due to inflation is permitted. Disposable income is therefore as in (29) with $l = 0$:

$$(60)\qquad d = (w + rk)(1 - \rho) - \pi m$$

while the government's budget equation is similarly

$$(61)\qquad \gamma y = (w + rk)\rho + (\pi + n)m$$

Using the firm's equilibrium conditions (4), (6) and the portfolio condition $m = \Lambda k$, the steady-state equations corresponding to the above model can be written

$$(62)\quad n(1 + \Lambda)k - \sigma[f(1 - \rho) - \pi\Lambda k] = 0$$

$$(63)\quad (\gamma - \rho)f - (\pi + n)\Lambda k = 0$$

Equations (62) and (63) determine the endogenous variables k and π.

We can now find the effects on these variables of changes in the tax rate ρ, and in

the government's budget proportion γ. Differentiating (62) and (63) totally, substituting for the equilibrium conditions, we get

$$\frac{dk}{d\rho} = 0 \tag{64}$$

$$\frac{d\pi}{d\rho} = \frac{-f}{\Lambda k} \tag{65}$$

The effects of changes in the government's budget proportion are similarly found:

$$\frac{dk}{d\gamma} = \frac{-\sigma f}{(f/k - f')\sigma(1-\gamma) + (1-\sigma)\pi\Lambda} \tag{66}$$

$$\frac{d\pi}{d\gamma} = \left[\Lambda^{-1}\frac{f}{k}\right. \left.(f/k - f')\sigma(1-\rho) - (1-\sigma)\pi\Lambda\right] \div [(f/k - f')\sigma(1-\gamma) + (1-\sigma)\pi\Lambda] \tag{67}$$

Under these conditions, a decreased tax rate can be fully compensated by a change in the rate of inflation without affecting the government's budget equation. This is because no substitution of real capital for money balances will take place under these extreme conditions. The full effect of tax changes falls on the rate of inflation.

On the other hand, expenditure changes necessitate a faster inflation rate. This lowers savings and hence capital intensity in the long run.

These results would basically not change if we allowed for only a partial inflationary loss offset to individuals or by the introduction of corporate taxes.

REFERENCES

M. Feldstein, "Inflation, Income Taxes and the Rate of Interest: A Theoretical Analysis," *Amer. Econ. Rev.*, Dec. 1976, *66*, 809–20.

Duncan Foley and Miguel Sidrauski, *Monetary and Fiscal Policy in a Growing Economy*, New York 1971.

D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.*, Sept. 1968, *58*, 713–53.

R. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Feb. 1956, *70*, 65–94.

Jerome Stein, *Money and Capacity Growth*, New York 1974.

J. Tobin, "Money and Economic Growth," *Econometrica*, Oct. 1965, *33*, 671–84.

Job Search, Labor Supply, and the Quit Decision: Theory and Evidence

By JOHN M. BARRON AND STEPHEN MCCAFFERTY*

Recently, several papers examined the quit decision. Donald Parsons considered a quit rate model based on the expected return to *employed* job search. Parsons relied on existing empirical evidence to justify the simplifying assumption in the model that workers quit only when a preferable job has been located. J. Peter Mattila provided further evidence that the majority of quits have lined up new jobs before quitting and suggested that quits into unemployment be viewed as "... a fairly small, constant exogenous flow ..." (p. 239).

In Section I, we provide a more complete theory of quit behavior within the context of an information and search approach. By identifying the cost of search as the utility value of time spent searching, new choice variables in optimal search strategy, the intensity of search and labor supply during search, are added.[1] This permits our model to encompass the *three* options facing an employed individual: employed job search, unemployed job search, or no job search. One result is that, contrary to the hypothesis of Mattila, the second option may be viewed as utility maximizing rather than "exogenous" behavior.

To test the theoretical predictions gained in Section I, a new economy wide measure of the quits entering unemployment, synonymous with the number choosing unemployed job search, is computed in Section II. One result, consistent with our model's prediction, is that quits entering unemployment are procyclical with the demand for labor. Section III contains concluding remarks.

I

This section sets up a model of the search behavior of an employed individual who seeks a higher wage. Let $f(w)$ denote the density function of wage offers and let v denote the probability that any search attempt results in a wage offer. We shall refer to v as the vacancy rate. If the individual sets a reservation (acceptance) wage of w^*, then the probability that a search attempt is successful (i.e., results in a wage offer greater than or equal to w^*) is given by $v \cdot k$ where

$$(1) \qquad k = \int_{w^*}^{\infty} f(w)dw$$

Assume that the only cost of search is the utility value of the time spent searching. If the individual can make α search attempts per unit of search time and the proportion of his time spent searching is τ, then the individual makes $\alpha \cdot \tau$ search attempts per unit time. Since the probability that a search attempt is successful is given by $v \cdot k$, the probability of a successful search attempt during a small time period ϵ is given by $\alpha \cdot \tau \cdot v \cdot k \cdot \epsilon$.

Taking the limit as ϵ approaches zero, the density function of the random variable T, the time at which search is successful, is exponential.[2] That is,

$$(2) \qquad h(T) = \alpha\tau\, vk(\exp(-\alpha\tau\, vkT))$$

During job search the individual attains a flow of utility from income and leisure given by

$$(3) \qquad U = U\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)$$

$$U_1, U_2 > 0; \quad U_{11}, U_{22} < 0; \quad U_{12} = 0$$

*Assistant professors of economics, Purdue University and Ohio State University, respectively. We wish to thank George Borts, John Carlson, John Kennan, J. Peter Mattila, Richard Peterson, and an anonymous referee for comments on an earlier draft.

[1]The model considers expected utility-maximizing behavior rather than expected income-maximizing behavior. A more complete discussion of search decisions in this context appears in McCafferty.

[2]See, for instance, John Freund (pp. 82; 111–12).

where y is the real value of nonwage income, w_0/p is his current real wage, and l is the fraction of time the individual decides to supply as labor. The arguments of U are constrained to be nonnegative. The individual's flow of utility on securing a better job rises to

$$(4) \qquad U = U\left(y + \frac{w}{p}\bar{l}, 1 - \bar{l}\right)$$

where w/p is the real wage the individual accepts (a random variable) and $\bar{l}$ is the fixed fraction of time comprising a new job's work week.[3]

Assume that the individual's perceptions of the wage offer distribution and the vacancy rate are in accord with reality. Then, from equations (3) and (4), the individual's expected level of utility with discount rate δ and planning horizon N is given by

$$(5) \qquad E(U) = E\left(\int_0^T U\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)e^{-\delta t}dt + \int_T^N U\left(y + \frac{w}{p}\bar{l}, 1 - \bar{l}\right)e^{-\delta t}dt\right)$$

with random variables T and w/p. The control variables in equation (5) are the labor supply l, the acceptance wage w^*, and search intensity τ. Unfortunately, the maximization of (5) poses serious technical difficulties. With a finite planning horizon, an optimum strategy will involve a decline in search intensity and the acceptance wage over time.[4] However, the density function $h(T)$ is derived under the assumption that the product $\alpha \cdot \tau \cdot \nu \cdot k$ is constant over time. If this product is not constant then the density function of T cannot be easily computed.

For most job searchers, however, the duration of search is small relative to the planning horizon and so these effects will be negligible. Therefore equation (5) can be reasonably approximated by the infinite horizon formulation

$$(6) \qquad E(U) = E\left(\int_0^T U\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)e^{-\delta t}dt + \int_T^\infty U\left(y + \frac{w}{p}\bar{l}, 1 - \bar{l}\right)e^{-\delta t}dt\right)$$

A recent paper by John Kennan shows that the extremum solution to equations (5) and (6) converge as N in equation (5) approaches infinity.

Taking the expectation of equation (6) with respect to T and w allows equation (6) to be rewritten as

$$(7) \qquad E(U) = \frac{1}{\delta}\left(U\left(y + \frac{w_0}{p}l, 1 - l - \tau\right) + \frac{\alpha\nu\tau k}{\delta + \alpha\nu\tau k}\int_{w^*}^\infty \left[U\left(y + \frac{w}{p}\bar{l}, 1 - \bar{l}\right) - U\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)\right]\frac{f(w)}{k(w^*)}dw\right)$$

The individual will seek to maximize expression (7) with respect to l, w^*, and τ. The first-order conditions for maximization are[5]

$$(8) \qquad \frac{w_0}{p} = \frac{U_2\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)}{U_1\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)}$$

$$(9) \qquad E(U) = U\left(y + \frac{w^*}{p}\bar{l}, 1 - \bar{l}\right)\Big/\delta$$

$$(10) \qquad U_2\left(y + \frac{w_0}{p}l, 1 - l - \tau\right) = \frac{\alpha\nu}{\delta + \alpha\tau}\int_{w^*}^\infty \left[U\left(y + \frac{w}{p}\bar{l}, 1 - \bar{l}\right) - U\left(y + \frac{w_0}{p}l, 1 - l - \tau\right)\right]f(w)dw$$

[3]The assumption that $\bar{l}$ is fixed is made only for convenience of exposition. If the new job had flexible hours the individual would just choose them to satisfy $w/p = U_2/U_1$. The crucial consideration is that the utility value of the job be monotonically related to its wage. This would still be the case.

[4]Reuben Gronau shows that the approach of the planning horizon will cause the acceptance wage to decline.

[5]The Appendix contains a more detailed derivation of (8), (9), and (10).

Expression (8) defines the conditions determining the rate of labor supply during search. Labor supply is chosen to equate the marginal utility of leisure with the marginal utility of the real wage. However, in the present context, leisure is net of both work and search.

The second expression, equation (9), generates the individual's acceptance wage. The acceptance wage should be that wage which leaves the individual indifferent between acceptance and continuing search with that acceptance wage. Thus the acceptance wage is chosen to equate the expected value of continued search ($E(U)$) with the value of accepting w^*.

The condition in equation (10) determines the individual's optimal search intensity τ. This condition shows that the individual should expand his search intensity until the marginal cost of search intensity expansion (the left-hand side term in equation (10)) equals the marginal expected gain to increased search intensity.

Differentiation of the first-order conditions generates the effects of exogenous disturbances on search behavior and labor supply during search. Below we consider the effects of changes in two exogenous variables, the vacancy rate and the current real wage.[6]

Results of a change in the vacancy rate on search intensity and labor supply are

$$(11)\quad \frac{d\tau}{dv} = \frac{-1}{v\left(1 + \frac{\alpha v}{\delta}\tau k\right)} \cdot \left(\frac{U_1^0}{U_{11}^0\left(\frac{w_0}{p}\right)} + \frac{1}{U_{22}^0}\right) > 0$$

$$(12)\quad \frac{dl}{dv} = \frac{1}{v\left(1 + \frac{\alpha v}{\delta}\tau k\right)} \cdot \frac{U_1^0}{U_{11}^0\left(\frac{w_0}{p}\right)} < 0$$

The superscript 0 denotes that the term is evaluated at the levels of consumption and leisure during search. We see that as the vacancy rate rises, the individual searches more and works less. This result is not surprising. Increases in the vacancy rate increase the returns to search, but do not affect the returns to work (the current real wage) or to leisure. Therefore there is a substitution away from both toward search.

In order to facilitate analysis of the effects of changes in the individual's current real wage, it is useful first to state explicitly the condition under which the standard labor supply curve is upward sloping. The effect of a change in (w_0/p) on the optimal value of l if τ and w^* are fixed can be obtained by differentiating equation (8) with respect to (w_0/p) and l. The result is that

$$(13)\quad \frac{\partial l}{\partial\left(\frac{w_0}{p}\right)}\bigg|_{\tau, w^*} = -\frac{U_1^0 + \frac{w_0}{p} l U_{11}^0}{U_{22}^0 + \left(\frac{w_0}{p}\right)^2 U_{11}^0}$$

The present analysis invokes the standard assumption of an upward sloping conventional labor supply curve, i.e., that equation (13) is positive.

Given that (13) is positive, results of a change in w_0/p on search intensity and labor supply are shown in equations (14) and (15). The first of these results, equation (14), shows that increases in the current real wage induce individuals to search less. There are two reasons for this result. First, since the real value of the individual's current wage has risen in relation to the distribution of wage offers, search becomes less productive and so the intensity of search declines. The second reason the intensity of search falls is related to the conventional assumption of an upward sloping labor supply curve. An increase in (w_0/p) induces the individual to work more and hence raises the marginal cost of search (foregone leisure).

Expression (15) shows that increases in the real wage increase labor supply during search. This results from the direct effect of the assumed upward sloping labor supply curve and the indirect effect of the allocation of time, formerly spent in search, to labor supply and leisure as search intensity falls.

[6]The Appendix contains a more detailed derivation of (11), (12), (14), and (15).

$$(14)\quad \frac{d\tau}{d\left(\frac{w_0}{p}\right)} = \frac{\frac{\alpha vkl}{\delta + \alpha v\tau k}\left(U_1^0 U_{22}^0 + U_1^0 U_{11}^0 \left(\frac{w_0}{p}\right)^2\right) + U_{22}^0\left(U_1^0 + U_{11}^0 l\frac{w_0}{p}\right)}{U_{11}^0 U_{22}^0\left(\frac{w_0}{p}\right)^2} < 0$$

$$(15)\quad \frac{dl}{d\left(\frac{w_0}{p}\right)} = \frac{-\left(U_1^0 + U_{11}^0 l\frac{w_0}{p}\right) - U_1^0 l\left(\frac{\alpha vk}{\delta + \alpha v\tau k}\right)}{U_{11}^0\left(\frac{w_0}{p}\right)^2} > 0$$

Of particular interest in the present discussion is the classification of workers into one of three groups. Either a worker doesn't search ($\tau = 0$), searches on the job ($\tau > 0$, $l > 0$), or quits to search ($l = 0$). From the first-order conditions, those who don't search (the contented workers) are those for whom the following are satisfied:

$$(16)\quad \frac{w_0}{p} = \frac{U_2\left(y + \frac{w_0}{p}l, 1 - l\right)}{U_1\left(y + \frac{w_0}{p}l, 1 - l\right)}$$

$$(17)\quad U_2\left(y + \frac{w_0}{p}l, 1 - l\right) \geq \frac{\alpha v}{\delta}\int_{w^*}^{\infty} \left[U\left(y + \frac{w}{p}\bar{l}, 1 - \bar{l}\right) - U\left(y + \frac{w_0}{p}l, 1 - l\right)\right] f(w)dw$$

Those who quit and search are those for whom the following is satisfied:

$$(18)\quad \frac{w_0}{p} < \frac{U_2(y, 1 - \tau)}{U_1(y, 1 - \tau)}$$

It is optimal to quit and search when the intensity of search is at a level that crowds out labor supply.

Let p_c denote the proportion of workers who do not search, p_e denote the proportion who search on the job, and let p_u denote the proportion of workers who quit to search. Consider now the effects on p_c, p_e, and p_u of an increase in the vacancy rate or a decrease in the current real wage among workers.

For some formerly contented workers, an increase in the vacancy rate will reverse inequality (17), and they will begin on-the-job search. The same result occurs for a decrease in the current real wage among workers. Thus,

$$(19)\quad p_c = p_c\left(v, \frac{w_0}{p}\right),\ \partial p_c/\partial v < 0,\ \partial p_c/\frac{\partial w_0}{p} > 0$$

An increase in the vacancy rate or a fall in (w_0/p) among workers results, according to equations (11)–(15), in a higher intensity of search and a lower labor supply of individuals currently engaged in on-the-job search. For some formerly employed job searchers, the rise in the intensity of search will be such that equation (18) now holds. Thus,

$$(20)\quad p_u = p_u\ v, \frac{w_0}{p}\ ,\ \partial p_u/\partial v > 0,\ \partial p_u/\partial\frac{\partial w_0}{p} < 0$$

The effect on p_e of an increase in vacancies or a decrease in current real wages is ambiguous. The fact that some previously contented workers start on-the-job search tend to increase the proportion of workers engaged in employed job search. However the resulting increased search intensity among workers also results in some fo

merly on-the-job searchers choosing unemployed job search.

Let q_e denote the quit rate for employed searchers who find a job and q_u denote the quit rate for those choosing unemployed job search. Then the following relations hold,

$$q_e = (\alpha \tau v k) p_e \tag{21}$$

$$q_u = p_u \tag{22}$$

where τ and k in equation (21) are averages over the on-the-job searchers.

Note that it is now no longer *theoretically* correct to assert that q_e is an increasing function of v, as Parsons has. Referring to equation (21), the reason is twofold. First, Parsons obtained an unambiguous increase in p_e by ignoring the option of unemployed search. From our model it is clear that $(p_e + p_u)$ increases, but it is not clear that p_e increases. Second, Parsons assumed a constant acceptance wage for an employed individual, such that the probability of finding an acceptable job $(\alpha \cdot \tau \cdot v \cdot k)$ unambiguously increases with an increase in v. Given transfer costs between jobs, a constant acceptance wage is consistent only with a series of job changes, the individual accepting any job paying a certain fraction above his current wage. This process of job changing continues until a wage is found high enough such that the individual becomes a contented worker. Our model seeks to emphasize costs that would lead the individual to economize on job changes, and thus views the individual as anticipating only one job change. In this case, the acceptance wage is an increasing function of v; thus, a rise in v and therefore τ would coincide with a fall in k and, without specifying particular functions, the net result for $\alpha \cdot \tau \cdot v \cdot k$ is uncertain.

Our model does provide unambiguous predictions on q_u. In particular, combining equations (19) and (22),

$$q_u = q_u\left(v, \frac{w_0}{p}\right),\ \partial q_u/\partial v > 0,\ \partial q_u/\partial \frac{w_0}{p} < 0 \tag{23}$$

To test equation (23), aggregate time-series data on quits into unemployment are computed in Section II. The data, however, require that cyclical effects on quits choosing unemployed job search be separated from effects of recent changes in labor force composition. One recent change in the composition of the labor force that appears to warrant consideration is the average age of employees. A discussion of this follows.

Over time, a particular worker accumulates training specific to a firm and training specific to an "occupation." Firms offer to workers some of the returns to each type of training to reduce the likelihood of quits. This observation is consistent with equations (20) and (23), suggesting that search intensity is inversely related to the current real wage. It follows that the levels of firm-specific and occupation-specific training, by directly affecting a nontransferable component of the current real wage, are inversely related to search intensity.

We shall contend that the level of each type of training depends directly on the length of employment in a particular job or occupation, and that these lengths are an increasing function of the age of a worker. Thus, equation (23) may be rewritten as

$$q_u = q_u(v, j(A)),\ \partial q_u/\partial A < 0 \tag{24}$$

where A is the average age of employees.

To capture a seasonal transfer cost postulated by Parsons, an August-September dummy variable, AS, is introduced. Given these considerations, the equation to be estimated and the predicted signs of the coefficients are

$$q_u = \exp(\alpha_0 + \alpha_1 AS) v^{\alpha_2} A^{\alpha_3} \tag{25}$$

$$\alpha_1 > 0,\ \alpha_2 > 0,\ \alpha_3 < 0$$

The exponential functional form chosen to estimate equation (24) follows Parsons.

II

Since the predictions of our model, specifically equation (25), conflict with previ-

ous work, it is important that a test of equation (25) be undertaken. Past empirical tests of quit behavior rely on manufacturing quit rates. However, these data reflect a measure of a total quit rate rather than a measure of q_u, the quit rate of those choosing unemployed job search. Fortunately, a measure of q_u, necessary for our purposes, may be obtained from available monthly survey data collected by the Bureau of Labor Statistics (*BLS*). The method of computing q_u follows.

Let r_t denote the flow of quits into unemployment in period t. Let G_t^{ab} denote the number unemployed through quits for at least a periods, but for no more than b periods in period t. If $a = 0, b - a = X$, and θ is the probability of remaining unemployed between periods for the interval $\{t, t - X\}$, then

$$G_t^{ab} = r_t + r_{t-1}\theta + r_{t-2}\theta^2 + \ldots + r_{t-X}\theta^X \tag{26}$$

Assuming a constant flow of quits r_t for the interval $\{t, t - X\}$, then

$$r_t = G_t^{ab}/(1 + \theta + \theta^2 + \ldots + \theta^X) \tag{27}$$

Let G_t^{bc} denote the number unemployed for at least b periods, but for not more than c periods in period t. Then, if $c - b = b - a$,

$$\theta = (G_t^{bc}/G_{t-X}^{ab})^{1/X} \tag{28}$$

We shall approximate G_t^{ab} and G_{t-X}^{ab} by the number of job leavers reported by the *BLS* to be unemployed less than five weeks in the survey week of months one $(t - X)$ and two (t). The G_t^{bc} is then approximated by the number of job leavers unemployed for at least five weeks but not more than ten weeks in the survey week of month two. Setting X equal to four and substituting these values into equations (27) and (28), an estimate of the weekly flow of quits into unemployment r_t is obtained. Since this estimate is centered at the end of month one, a measure of the monthly quit rate into unemployment, representing q_u, is given for month two by[7]

$$q_u = 2(r_t + r_{t+X})/e_t \tag{29}$$

where e_t is the total number employed in month two.

The results of estimating equation (25) using the monthly values of q_u so computed and approximating the vacancy rate by the help wanted advertising index compiled by The National Industrial Conference Board are

$$\begin{aligned} \ln q_u = \; & \underset{(1.75)}{10.82} + \underset{(2.53)}{.486} \ln v \\ & - \underset{(2.27)}{3.77} \ln A + \underset{(6.55)}{.212} AS \end{aligned} \tag{30}$$

$$R^2 = .76$$

Given data limitations, the period of observations is from February 1967 to August 1975. Absolute values of the t-statistics appear in parentheses.[8]

For the period 1967 to 1975, the average age of the labor force A has exhibited a strong downward trend. This raises questions as to whether A serves as a proxy for other neglected variables correlated with time. To better determine the potential effects of age on q_u, a cross-section sample is thus required. Though the available data is limited, we were able to obtain measures of quit rates into unemployment for individuals aged 16 to 19 and for individuals 20 years of age and older. It was found that the quit rate into unemployment for those aged 16 to 19 averaged four and one-half times the quit rate into unemployment for individuals over 20. This is consistent with our hypothesized effect of age on q_u and with the sign and significance of the coefficient on A in equation (30). The size of the coefficient on A must still, however, be viewed with some caution.

Evidence of the cyclical behavior of q_u with respect to total quits is possible if we

[7]Since those unemployed between five and ten weeks in month two include some not recorded as unemployed less than five weeks in month one, the actual estimate of θ has an upward bias. The actual estimate of the quit rate into unemployment thus has a downward bias.

[8]Equation (30) is corrected for first-order serially correlated errors. The rho value is .677.

assume the published manufacturing quit rate q_m measures the economy-wide total quit rate. Then q_u/q_m represents the proportion of quits that enter unemployment. The relation between q_u/q_m and the vacancy rate is given by[9]

$$(31)\quad ln(q_u/q_m) = \underset{(1.53)}{1.79} - \underset{(2.87)}{.729}\, ln\, v \qquad R^2 = .616$$

III

Several findings emerge from this study. One is that, on both theoretical and empirical grounds, the quit rate into unemployment q_u is responsive to the vacancy rate. As noted before, this finding differs from the predictions concerning q_u made by Mattila. However, Mattila's point that most quitters do not quit to search is supported. We found that during the period the proportion of quits entering unemployment q_u/q_m, ranged from .14 to .42 with a mean of .21.

A second finding is that the elasticity of 49 for the quit rate into unemployment with respect to the vacancy rate is well below the quit rate (total)–vacancy elasticity range of 1.0–2.0 discovered by Parsons. Thus, a fall in the vacancy rate increases the proportion of quits entering unemployment. Further evidence supporting this cyclical behavior of q_u with respect to the total quit rate is provided by equation (31).

A third finding is that the recent 4 percent reduction in the average age of workers from 1967 to 1975 implies, *ceteris paribus*, a 15 percent increase in the quit rate into unemployment.

APPENDIX

Adopting the convention that $c \equiv \alpha v/\delta$, equation (7) may be rewritten as

$$(A1)\quad V \equiv \delta E(U) = U\left(y + \frac{w_0}{p} l, 1 - l - \tau\right) + \frac{c\tau k}{1 + c\tau k} \int_{w^*}^{\infty} \Delta U \frac{f(w)}{k(w^*)}\, dw$$

where

$$\Delta U \equiv U\left(y + \frac{w}{p} \bar{l}, 1 - \bar{l}\right) - U\left(y + \frac{w_0}{p} l, 1 - l - \tau\right)$$

Differentiation of equation (A1) permits the derivation of the first-order conditions

$$(A2)\quad \frac{\partial V}{\partial \tau} = \frac{1}{1 + c\tau k}\left[-U_2^0 + \frac{ck}{1 + c\tau k} \int_{w^*}^{\infty} \Delta U \frac{f(w)}{k(w^*)}\, dw\right] = 0$$

$$(A3)\quad \frac{\partial V}{\partial w^*} = \frac{c\tau f(w^*)}{1 + c\tau k}\left[\frac{c\tau k}{1 + c\tau k} \int_{w^*}^{\infty} \Delta U \frac{f(w)}{k(w^*)}\, dw - \Delta U^*\right] = 0$$

where

$$\Delta U^* \equiv U\left(y + \frac{w^*}{p} \bar{l}, 1 - \bar{l}\right) - U\left(y + \frac{w_0}{p}, 1 - l - \tau\right)$$

and

$$(A4)\quad \frac{\partial V}{\partial l} = \frac{1}{1 + c\tau k}\left[\frac{w_0}{p} U_1^0 - U_2^0\right] = 0$$

The superscript 0 again means that the utility function is evaluated at the levels of consumption and leisure during search. Equations (A2) and (A4) correspond directly with equations (10) and (8), respectively. Equation (9) can be derived by combining equations (A1) and (A3).

By differentiating the first-order conditions, we can derive the second-order matrix. The computations are as follows: first differentiating (A2) with respect to τ yields

$$(A5)\quad \frac{\partial^2 V}{\partial \tau^2} = \frac{1}{1 + c\tau k}\left[U_{22}^0 + \frac{ck}{1 + c\tau k}\left[\frac{2ck}{1 + c\tau k}\left[U_2^0 - \frac{ck}{1 + c\tau k} \int_{w^*}^{\infty} \Delta U \frac{f(w)}{k(w^*)}\, dw\right]\right]\right]$$

[9] Equation (31) is corrected for first-order serially correlated errors. The rho value is .689.

Combining equations (A5) and (A2) we note that

$$\text{(A6)} \qquad \frac{\partial^2 V}{\partial \tau^2} = \frac{U^0_{22}}{1 + c\tau k}$$

Differentiating equation (A3) with respect to w^* yields

$$\text{(A7)} \quad \frac{\partial^2 V}{\partial w^{*2}} = \frac{c\tau f(w^*)}{1 + c\tau k}\left[\left[\frac{2(c\tau)^2 f(w^*)}{(1 + c\tau k)^2} \int_{w^*}^{\infty} \Delta U \frac{f(w)}{k(w^*)} dw \frac{c\tau f(w^*)}{1 + c\tau k} \Delta U^*\right] - \frac{\bar{l}}{p} U^*_1\right] + \frac{f'(w^*)}{f(w^*)}\left[\frac{\partial V}{\partial w^*}\right]$$

Now combining equations (A7) and (A3) permits the derivation of

$$\text{(A8)} \quad \frac{\partial^2 V}{\partial w^{*2}} = -\frac{c\tau f(w^*)}{1 + c\tau k}\frac{\bar{l}}{p} U^*_1$$

Differentiating equation (A4) with respect to l, and recalling the assumption that $U_{12} = 0$ yields

$$\text{(A9)} \quad \frac{\partial^2 V}{\partial l^2} = \frac{1}{1 + c\tau k}\left[\left(\frac{w_0}{p}\right)^2 U^0_{11} + U^0_{22}\right]$$

Differentiation of equation (A2) with respect to w^* yields

$$\text{(A10)} \quad \frac{\partial^2 V}{\partial \tau \partial w^*} = \frac{1}{1 + c\tau k}\left[\frac{c^2 \tau f(w^*)}{(1 + c\tau k)^2} \int_{w^*}^{\infty} \Delta U f(w) dw - \frac{c f(w^*)}{1 + c\tau k} \Delta U^*\right]$$

Combining equations (A10) and (A3) and the fact that $V_{ij} = V_{ji}$ permits derivation of

$$\text{(A11)} \qquad \frac{\partial^2 V}{\partial \tau \partial w^*} = \frac{\partial^2 V}{\partial w^* \partial \tau} = 0$$

Recalling that $U_{12} = 0$, differentiation of equation (A2) with respect to l yields

$$\text{(A12)} \quad \frac{\partial^2 V}{\partial \tau \partial l} = \frac{1}{1 + c\tau k}\left[U^0_{22} - \frac{ck}{1 + c\tau k}\left[\frac{w_0}{p} U^0_1 - U^0_2\right]\right]$$

Now combining equations (A12) and (A4) we find that

$$\text{(A13)} \quad \frac{\partial^2 V}{\partial \tau \partial l} = \frac{\partial^2 V}{\partial l \partial \tau} = \frac{U^0_{22}}{1 + c\tau k}$$

Finally, differentiating equation (A4) with respect to w^* yields

$$\text{(A14)} \quad \frac{\partial^2 V}{\partial l \partial w^*} = \frac{\partial^2 V}{\partial w^* \partial l} = 0$$

The second-order matrix for the maximization of equation (6) can now be written as A(15).

$$\text{(A15)} \quad \begin{bmatrix} \frac{\partial^2 V}{\partial l^2} & \frac{\partial^2 V}{\partial l \partial w^*} & \frac{\partial^2 V}{\partial l \partial \tau} \\ \frac{\partial^2 V}{\partial w^* \partial l} & \frac{\partial^2 V}{\partial w^{*2}} & \frac{\partial^2 V}{\partial w^* \partial \tau} \\ \frac{\partial^2 V}{\partial \tau \partial l} & \frac{\partial^2 V}{\partial \tau \partial w^*} & \frac{\partial^2 V}{\partial \tau^2} \end{bmatrix} = \begin{bmatrix} \frac{U^0_{22} + \left(\frac{w_0}{p}\right)^2 U^0_{11}}{1 + c\tau k} & 0 & \frac{U^0_{22}}{1 + c\tau k} \\ 0 & -\frac{c\tau f(w^*)\bar{l}/p\, U^*_1}{1 + c\tau k} & 0 \\ \frac{U^0_{22}}{1 + c\tau k} & 0 & \frac{U^0_{22}}{1 + c\tau k} \end{bmatrix}$$

Straightforward examination of equation (A15) assures that the second-order condi-

---

$$\text{(A16)} \quad \begin{bmatrix} U^0_{22} + \left(\frac{w_0}{p}\right)^2 U^0_{11} & 0 & U^0_{22} \\ 0 & \bar{l}/p\, U^*_1 & 0 \\ U^0_{22} & 0 & U^0_{22} \end{bmatrix} \begin{bmatrix} dl \\ dw^* \\ d\tau \end{bmatrix} = \begin{bmatrix} 0 & -U^0_1 \frac{w_0}{p} + lU^0_{11} \\ \frac{\tau U^0_2}{1 + c\tau k} & \frac{lU^0_1}{1 + c\tau k} \\ \frac{U^0_2}{1 + c\tau k} & \frac{cklU^0_1}{1 + c\tau k} \end{bmatrix} \begin{bmatrix} dc \\ d\left(\frac{w_0}{p}\right) \end{bmatrix}$$

tions for the maximization of equation (6) are unambiguously satisfied.

Now differentiating equations (A2), (A3), and (A4) with respect to τ, $\dot{w}^*$, l, c, and w_0/p, one may derive A(16). Application of Cramer's rule facilitates the derivations of equations (11), (12), (14), and (15) in the text.

REFERENCES

Gary S. Becker, *Human Capital,* New York 1964.

J. P. Danforth, "Expected Utility, Mandatory Retirement and Job Search," Center Econ. Res., disc. pap. 74-41, Univ. Minnesota, June 1974.

John E. Freund, *Mathematical Statistics,* Englewood Cliffs 1971.

R. Gronau, "Information and Frictional Unemployment," *Amer. Econ. Rev.,* June 1971, *61*, 290-301.

J. Kennan, "Expected Utility Maximization: Some General Convergence Results," unpublished paper, Brown Univ. 1975.

J. P. Mattila, "Job Quitting and Frictional Unemployment," *Amer. Econ. Rev.,* Mar. 1974, *64*, 235-39.

S. McCafferty, "A Theory of Search for Transaction Partners," unpublished doctoral dissertation, Brown Univ. 1977.

D. O. Parson, "Quit Rates Over Time: A Search and Information Approach," *Amer. Econ. Rev.,* June 1973, *63*, 390-401.

J. Seater, "A Unified Model of Consumption, Labor Supply and Job Search," unpublished doctoral dissertation, Brown Univ. 1974.

U.S. Bureau of Labor Statistics, *Employment and Earnings,* various years.

———, unpublished data.

U.S. Department of Commerce, *Business Conditions Digest,* various years.

Faculty Skills and the Salary Structure in Academe: A Market Perspective

By HOWARD P. TUCKMAN, JAMES H. GAPINSKI, AND ROBERT P. HAGEMANN*

> In some universities the salary makes but a part, and frequently but a small part of the emoluments of the teacher, of which the greater part arises from the honoraries or fees of his pupils. The necessity of application, though always more or less diminished, is not in this case entirely taken away. Reputation in his profession is still of some importance to him, and he still has some dependency upon the affection, gratitude, and favourable report of those who have attended upon his instruction; and these favourable sentiments he is likely to gain in no way so well as by deserving them, that is, by the abilities and diligence with which he discharges every part of his duty.
>
> Adam Smith, *The Wealth of Nations*, 1776

A number of academicians have recently addressed a pragmatic question which strikes very close to their pocketbook. Stripped of its ornamentation, it reads, "What determines my salary?" These researchers, following their own theoretical, methodological, and professional predilections, have formulated and tested several different salary equations. They have also examined various subissues, among them whether salary differentials exist by sex. This paper brings to the inquiry a new comprehensive data set, and it hypothesizes a relationship between salaries and the functioning of the academic marketplace. Specifically, we postulate that the salaries of faculty members are determined to a substantial degree by market valuation of their skills. These skill markets may differ by discipline and sex, and the rewards for given skills may differ accordingly. This occasions several implications ignored in the earlier studies of David Katz, Emily Hoffman, and others. First, to the extent that discrimination exists in academe, it may cause a differential not only in the average salary of male and female faculty but also in the returns to specific skills. A corollary proposition is that schemes designed to equalize average salaries may create new inequities and lead to a misallocation of resources for reasons suggested elsewhere by Tuckman. Second, to the extent that the returns to a given skill differ by discipline, statements of the return to that skill for all fields combined provide a misleading impression of the reward enjoyed by faculty in any particular field. Concomitantly, union efforts to negotiate a single salary schedule for all faculty will have different consequences for individual faculty, both in terms of their average salary levels and in terms of the returns they receive for specific skills. If the supply of these skills is price responsive, then the effect of unionization on that supply, measured relative to the supply in the absence of unionization, will differ by field. Third, to the extent that faculty react to differences in the returns to individual skills, administrative actions aimed at changing faculty behavior which ignore such differences are likely to be at best only partially successful. For these and related reasons, we believe that it is analytically useful to introduce the concept of a market for faculty skills, acknowledging our debt to the human capital antecessor.

I. The Market for Faculty Skills

Many academic departments desire faculty skilled in teaching, research, public service, and administration. Individuals with these skills can provide benefits to

*Associate professors of economics and doctoral candidate, respectively, Florida State University. We sincerely thank Alan E. Bayer for providing us with the data used in this study. The research was made possible by an NSF-RANN grant, SSH72-03432 A02 (formerly GI-34394).

their department in the form of increased student enrollments, outside grant funding, and recognition by the university, local community, and discipline at large. Two points seem evident. First, faculty members usually do not possess these skills in equal measure.[1] Effective administrators are not necessarily the most able researchers, and those skilled in public service may be weak teachers. Since skills take time to develop, faculty to some extent choose among competing alternatives, and the return from each skill may affect their decisions. Moreover, the diverse returns may have influenced the choice skills developed in graduate school. Second, departments face an allocation problem. Given limited resources, they can hire only a few faculty. Whom a department selects depends on the importance it gives to the "package" of skills offered by each potential faculty member and on the premium those skills command in the market.[2] Tenure regulations which limit a department's alternatives prohibit rapid adaptation to changing departmental needs.

The stock of each faculty skill available in the marketplace is relatively fixed in the short run. If the demand for a given skill increases, the price paid to faculty for this skill tends to increase, creating salary differentials among faculty. In the long run, the number of faculty possessing the desired skill also increases, thus narrowing the differential among skills.[3]

Is there a market return to outstanding teaching? The evidence is mixed. In the absence of a comprehensive theory of learning, departments may not effectively appraise the teaching output of their faculty. Nevertheless, the inquisitive department head or other decision maker can probably identify outstanding and poor teachers. If a department wants to offer a salary increment to good teachers, it can discriminate among faculty on the basis of their teaching abilities. A problem arises when outstanding teachers attempt to sell their skills to other institutions or when new Ph.D.'s with untested skills enter the job market. Good teachers are normally known locally; it is difficult to gain a national reputation for one's teaching skills as David Brown suggests (1965, pp. 203–06). Consequently, the demand for outstanding teaching skills may be limited and the price paid for these skills low.

The output of researchers is more visible, consisting of articles, books, and other published pieces which often attract a national audience. The quality of a researcher's work can be more readily judged by experts; its worth evaluated in terms of the grants it brings, its effects on the national reputation of the researcher.[4] Since researchers may also be more versatile with quantitative and analytic techniques than are teachers, and thus more substitutable in other occupations, both supply and demand forces would seem to place a relative premium on research skills. This should be at least partially reflected in the salary increments received from publication.

Public service involves work with communities and public organizations, with departmental or university committees, and with charitable or educational organizations. Some departments regard these activities highly and demand faculty be successful in them. However, as in the case of teaching, such activities are more inclined to receive local rather than national recognition. The market for faculty with these

[1]Alternatively, it might be argued that faculty possess these skills in equal measure but fail to cultivate them equally. For a polemic explanation of why this occurs, see Pierre van den Berghe, ch. 6.

[2]The premiums for new faculty are lower than those for established faculty since the latter are more likely to have invested time in a specialty. Once a person has opted to cultivate a particular skill, it is difficult to change direction since time is necessary to make a shift. This introduces short-run rigidities into the marketplace.

[3]Salary need not be the only factor causing faculty to augment their background in the demanded skills. Greater job options, increased prestige, and opportunities for creative work also play a role. Nothing reported in this paper is intended to minimize the importance of nonpecuniary returns.

[4]Not all research has practical value. Moreover, the significance of seminal work is often recognized by hindsight. Some types of research activities are of immediate value to a department, and these are most apt to be rewarded. Whether such reward policies impede in-depth scientific inquiry has not yet been analyzed.

skills may be circumscribed given the difficulties inherent in determining a faculty member's public service abilities.

Administrative skills are largely learned on the job. While grant management, departmental and university duties, and prior work experiences provide faculty with some skills, much administrative experience is human capital specific. Furthermore, administrative skills are not easily measured, and thus the market for this type of skill may be limited.[5]

Most faculty enter the job market possessing more than one skill, and the salary an individual is offered presumably includes a return for each skill valued by the employing department. Since different disciplines may assign different weights to given skills, the structure of salaries is likely to vary by discipline. It may also vary by sex as males and females may be subject to dissimilar supply and demand phenomena as suggested by Barbara Reagan.[6]

II. The Data Base and Model Specifications

This paper employs data gathered by the American Council on Education (*ACE*) as part of a 1972–73 national cross-section study of faculty. The *ACE* initially selected 301 institutions representing diverse institutional types, levels of selectivity, and amounts of institutional wealth. Included were 78 universities, 181 four-year colleges, and 42 junior or community colleges.[7] However, our interest in the returns to various academic skills dictated that of all respondents to the *ACE* questionnaire only full-time university faculty be considered. These are drawn from five disciplines grouped according to a widely accepted classification scheme: *Social Sciences*—Anthropology, Geography, Political Science, Sociology, Economics, History, Psychology; *Liberal Arts*—English, Music; *Math-Engineering*—Civil Engineering, Electrical Engineering, Mathematics; *Biological Sciences*—Biochemistry, Botany, Zoology; *Physical Sciences* — Chemistry, Earth Science, Physics. The revised data file, with incomplete responses deleted, consists of 12,685 faculty of which 11,973 are male and 712 are female.[8]

To determine the returns to select skills, a model is postulated which has roots in those of Katz, and George Johnson and Frank Stafford. A list and brief discussion of the variables follow.

Salary: Income received by a faculty member from the employing institution for contractual services. Excluded are consulting fees, royalties, and other income earned outside the institution, the implicit assumption being that universities do not consider a person's outside source of income in setting salary levels.

Articles and Books: Total published as of 1972–73. These variables are taken as proxies for research skill. The number of articles is partitioned into six groups (1–2, 3–4, 5–10, 11–20, 21–50, >50), each represented by a dummy variable which assumes a unit value when articles published fall in the corresponding group and zero otherwise. Books are partitioned into four categories (1–2, 3–4, 5–10, and >10) with a dummy assigned to each. The same "1-0" criterion applies. This partitioning of publications can be used to examine whether they have a linear or non-linear effect on

[5]Administrators usually (but not always) earn at least as much as the most highly paid people they supervise. Since the salaries faculty receive vary, the average administrator earns more than the average nonadministrator. Whether this constitutes a return to the administrative skill, to longer hours, to foregone alternatives, or the prerogatives of office remains to be established.

[6]For a general presentation of this view, see Brown (1967, pp. 62–63,66). But note that Brown does not develop the implications of this argument in the context of investment in specific skills.

[7]The original mailing involved 108,722 individuals; two following mailings together with the original, produced 53,034 responses. For more on the sample, see Alan Bayer, pp. 1–5.

[8]The *ACE* developed a complex weighting scheme for making the faculty data representative of the entire population of college and university teaching faculty. These have not been used here since the weights are based on the teaching rather than the total faculty community. The breadth of our sample, however, suggests that it should be at least as representative as the NSF National Register.

faculty salaries. The *ACE* did not compile continuous publication variables.

Teaching Award: A dummy variable with unity indicating that an individual has received a teaching award. While a measure of teaching quality ranging from poor to outstanding might be preferable to the award variable, such a measure is unavailable as some departments sampled do not rate faculty on teaching performance and others employ disparate rating schemes. It is probable, however, that our award variable identifies a high proportion of those with outstanding teaching skills.[9]

Public Service: A dummy variable which equals unity if the faculty member is currently engaged in unpaid public service. Since organizational work typically involved long-term commitments, the current service variable is assumed to reflect past service as well.

Administration: This skill is introduced by two dummies. The first denotes (by 1) the individual who currently lists administration as the prime work activity; the second denotes (by 1) the faculty member who previously was a dean or department head.[10]

Experience: Number of years since the person received the highest degree. It enters as a quadratic to allow for diminishing marginal returns to experience. Since skill variables are explicitly included in the model, this variable may be interpreted as measuring the effect of experience on salary net of those increments which result from the cultivation of the specific skill.

Ph.D.: A dummy variable with a unit value when a person has a Ph.D. or its equivalent.

[9]In fact, the variable may identify too many outstanding teachers: almost 20 percent of the faculty shown in Table 1 received an outstanding teaching award This proportion seems very high, although it is difficult to find a yardstick against which it can be judged

[10]The dummy variables for administration and public service stand for activities rather than skills. It seems reasonable to assume, however, that faculty participating in these endeavors inherently possess a certain level of the requisite skills and enhance them in the line of duty.

Start: Defined as year of highest degree minus year of birth, this variable represents starting age at the point at which the highest degree is received. It enters interactively with the Ph.D. and experience variables.

Eleven-Month Salary and *Quality of Department:* Dummy variables the first of which assumes unity if the contractual period of employment is 11 months. Quality enters through two dummies with unity assigned to the relevant variable if the department's rating falls in the 3.1–4.0 or 4.1–5.0 interval of the Roose-Anderson scale. The most favorable rating is 5.

Region of Department: Dummy variables from North, Great Lakes, and Southeast to allow for regional differences in labor markets. Unity is assigned to a dummy when a department is located in the corresponding region. Southwest and West combine to form the region of reference.

Black: A dummy with a unit value when the faculty member is black.

An ordinary least squares equation was fitted to the *ACE* data and a complete analysis of covariance conducted to test whether the salary structure could be regarded as identical across the ten discipline-sex groups. The answer was unequivocally negative; the null hypotheses of overall, intercept, and slope homogeneity were all soundly rejected at the 5 percent level. But perhaps the salary structure differed only by sex, not by discipline; that is, males might display one uniform structure across disciplines and females another. The three hypotheses for this scenario were tested at 5 percent and rejected. A third possibility of homogeneity across sexes for given disciplines remained. This posed a minor problem because of the small sample sizes for females in liberal arts, math-engineering, biological sciences, and physical sciences. The female regressions for these four disciplines seemed unrepresentative, and consequently these data groups were deleted from further consideration. For social sciences, which boasted a much larger female sample, the three hypotheses

TABLE 1—ESTIMATED SALARY EQUATION FOR MALE FACULTY BY DISCIPLINE
(Shown in dollars)

	Regression Coefficients				
	Soc Sci	Lib Art	Math-Eng	Bio Sci	Phy Sci
Articles					
1-2	428[d]	318	1,040[b]	260	1,296
3-4	562[c]	780[c]	1,383[b]	-175	1,489[d]
5-10	1,010[b]	1,224[b]	1,558[b]	56	1,488[c]
11-20	1,758[b]	2,387[b]	2,772[b]	1,032	1,837[b]
21-50	2,621[b]	3,435[b]	4,352[b]	1,982	3,170[b]
>50	4,219[b]	5,851[b]	6,205[b]	5,006[b]	6,016[b]
Books					
1-2	426[b]	122	444[c]	-426	432[c]
3-4	1,424[b]	777[c]	1,073[b]	639	625[c]
5-10	2,515[b]	1,710[b]	1,407[b]	515	904[c]
>10	1,995[b]	2,216[b]	-441	699	1,067
Teaching Award	276[d]	280	217	615[d]	-246
Public Service	643[b]	410[d]	633[b]	175	675[b]
Administration					
Current	3,403[b]	2,988[b]	2,605[b]	3,809[b]	3,610[b]
Previous	1,448[b]	2,337[b]	1,964[b]	1,811[b]	1,943[b]
Experience	528[b]	279[b]	588[b]	546[b]	547[b]
Experience Squared	-8[b]	-2[d]	-8[b]	-8[b]	-9[b]
Ph.D.[a]	1,925[c]	-1,938[c]	2,716[b]	-527	1,852[d]
Start	43[d]	-83[b]	81[b]	-13	33
Start x Experience	-3	-.4	-3[b]	-3[c]	4
Start x Ph.D.	-26	114[b]	-16	88	-23
Quality of Department					
3.1-4.0	1,098[b]	841[b]	743[b]	2,185[b]	646[b]
4.1-5.0	1,297[b]	842	1,030[b]	2,016[b]	1,539[b]
Region of Department					
North	1,426[b]	700[c]	1,652[b]	1,089[b]	872[b]
Great Lakes	844[b]	621[c]	787[b]	-177	657[b]
Southeast	1,187[b]	372	1,133[b]	341	837[b]
Eleven-Month Salary	2,784[b]	1,214[b]	3,537[b]	2,608[b]	3,520[b]
Black Faculty	858	120	171	1,056	1,403
Constant	6,543[b]	10,781[b]	4,416[b]	7,865[b]	5,335[b]
R^2	.59	.54	.62	.62	.61
Sample Size	4,687	1,497	2,195	1,046	2,548

[a]The negative entries in the Ph.D. row do not translate into negative returns to the Ph.D. for reasons explained in the text.
[b]Significant at 1 percent.
[c]Significant at 5 percent.
[d]Significant at 10 percent.

were rejected at 5 percent. This evidence urged a separate investigation of salary for each discipline-sex group.

III. Empirical Analysis of the Salary Structure for Males

Table 1 presents the results obtained by fitting the salary equation to data on male faculty in each discipline. The coefficient for any variable in a given set of dummies is interpreted relative to the category excluded from that set. For example, the coefficient of 3-4 articles shows the extra salary which a faculty having that number of articles earns relative to one who published no articles.[11] All significance tests are F's. To

[11]A skill coefficient reflects the "mean circumstance" of the faculty already in that category. It might not signify the return which a person just entering that category would receive.

illustrate the computations of salary from the above regression we have selected a social scientist with 1-2 articles, 1-2 books, no teaching award and no public service or administrative background, 10 years of experience, a Ph.D. obtained at age 31, in a non-"quality" northern department. The relevant computation, with the coefficients of the variables shown in parentheses, is as follows: ($428) + ($426) + ($528)10 − ($8)10 · 10 + ($1925) + ($43)31 − ($.30)31 · 10 − ($26)31 + ($1426) + ($6543) = $15,662.

Salaries of males who publish articles generally rise monotonically with articles produced. A departure from this pattern occurs for the biological sciences in the first few categories. The configurations for book coefficients are considerably more disparate across fields. The liberal arts and physical sciences coefficients suggest continual increases; those for social sciences and math-engineering trace an inverted V, the drop being associated with category >10. The negative coefficient for math-engineering is insignificant; the corresponding coefficient for social sciences is significant and may reflect a tendency for the highest category to contain revised texts, readers, or other edited volumes not valued by departments in that discipline. Unfortunately, the *ACE* data are not complete enough to allow a test of this hypothesis. The book coefficients for biological sciences display a saw-toothed movement, but none are significant.

From the article and book coefficients, a return per publication can be extracted. Those figures appear in Table 2. In computing average returns, the midpoint of each publication interval was used; midpoints for >50 and >10 were taken as 65 and 12, respectively. Except in the maverick biological sciences, the articles evidence a tendency for diminishing returns, the rate of decline for the first several articles differing strikingly by field. Average returns for books either decline continually or first rise then fall.

Outstanding teaching appears to yield a low rate of return; in four of the five disciplines teaching excellence involves a smaller

TABLE 2—RETURN PER PUBLICATION BY MALE FACULTY IN EACH DISCIPLINE

	Soc Sci	Lib Art	Math-Eng	Bio Sci	Phy Sci
Articles					
1-2	$285	$212	$693	$173	$864
3-4	161	223	395	−50	425
5-10	135	163	208	7	198
11-20	113	154	179	67	119
21-50	74	97	123	56	89
>50	65	90	95	77	93
Books					
1-2	284	81	296	−284	288
3-4	407	222	307	183	179
5-10	335	228	188	69	121
>10	166	185	−37	58	89

return than even nominal publication of articles. Public service generally seems more lucrative than teaching, but those with 3-4 articles receive a higher return.[12] Current administrators enjoy larger salary adjustments, earning at least $2,600 more than those not engaged in this activity. In each discipline the gain from current administration is at minimum three times that from teaching and public service combined. It is matched only by extensive article publication.

The experience variables suggest the existence of diminishing marginal returns in all disciplines, a finding supported by other studies.[13] The severity of diminishing returns differs across fields, however, as does the point of negative returns. Figure 1 illustrates this point. Diminishing returns set in most gradually for liberal arts, and there negative returns do not materialize through 35 years of experience. For the

[12] Although these categories are not mutually exclusive, the time constraint imposes a tradeoff on faculty, at least to some extent. Thus, this comparison seems warranted

[13] These diminishing returns may reflect decay in the average faculty member's stock of human capital in each field, obsolescence of knowledge as new knowledge is introduced, generational or cohort effects, or consequences of institutional practices. These factors are difficult to disentangle empirically. It should be noted that the experience-earnings profiles presented in this article do not conform to the conventional wisdom as to which fields evidence the most rapid obsolescence of knowledge.

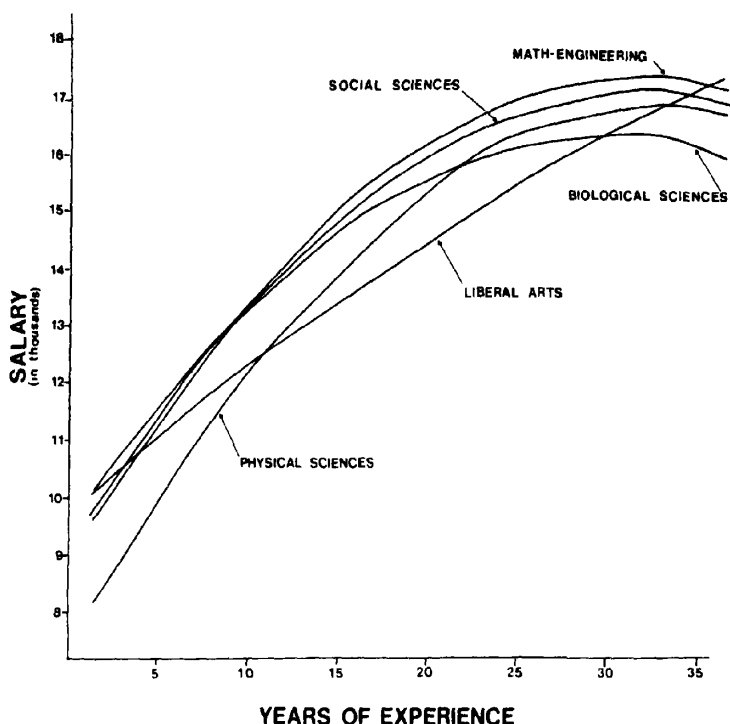



FIGURE 1 EXPERIENCE-EARNINGS PROFILE FOR MALE FACULTY BY DISCIPLINE

other disciplines, negative returns surface between 30 and 35 years of experience.[14] The individuals being compared are whites who completed their doctorates at age 31 and who are employed on a nine-month basis in an unranked western department. Publishers, outstanding teachers, public servants, and administrators are ruled out.

In conformity with casual empiricism, the Ph.D. always increases salary. This direct relationship is masked at times, but it readily emerges when the interactive component with Start is taken into account. For example, given a starting age of 31, the Ph.D. coefficients for liberal arts and biological sciences become $1,543 and $2,175, respectively. Black faculty are found to earn more than whites, but the regression coefficients are all insignificant. The small number of faculty in this category precluded more elaborate analysis.

IV. Contrapuntal Structures For Women and Men

The salary equation for female social scientists is reported in Table 3. Article publication is associated with a monotonic salary improvement save in the extreme category, where the coefficient is supported by a cell count of only 3. This pattern for articles bears a noticeable similarity to that for male social scientists. The configuration of book coefficients resembles the men's, and the negative coefficient for >10 category reflects the predicament of two women. Current administrators again receive handsome returns, though less than their male counterparts.

[14]As scholars mature their list of accomplishments grows, and the resulting salary increments impart "steps" to the experience-earnings loci. This complication has been omitted from Figure 1 under *ceteris paribus*.

TABLE 3—ESTIMATED SALARY EQUATION AND RETURN PER PUBLICATION FOR FEMALE FACULTY IN SOCIAL SCIENCES
(Shown in dollars)

	Regression Coefficient	Average Return
Articles		
1-2	212	141
3-4	728	208
5-10	877[c]	117
11-20	1,168[b]	75
21-50	2,434[a]	69
>50	200	3
Books		
1-2	355	237
3-4	1,323[a]	378
5-10	2,044[a]	273
>10	-130	-11
Teaching Award	418	
Public Service	722[a]	
Administration		
Current	1,530[a]	
Previous	1,956[a]	
Experience	446[a]	
Experience Squared	-7[a]	
Ph D.	4,427[a]	
Start	94[c]	
Start x Experience	-2	
Start x Ph D.	-65	
Quality of Department		
3.1-4.0	455	
4.1-5.0	382	
Region of Department		
North	1,443[a]	
Great Lakes	1,076[a]	
Southeast	1,609[a]	
Eleven-Month Salary	2,168[a]	
Black Faculty	1,987[c]	
Constant	3,546[c]	
R^2	.58	
Sample Size	371	

[a]Significant at 1 percent.
[b]Significant at 5 percent.
[c]Significant at 10 percent.

Experience affects salaries in the manner suggested by Johnson and Stafford. The coefficient of the experience variable, including the interaction term with Start, is less positive for females than for males while that of the quadratic term is less negative for females. A catch-up phenomenon is thus implied; the salary differential although increasing initially increases at a decreasing rate until it begins to shrink. Given sufficient time, the differential would be driven to zero. Figure 2 shows the experience-earnings profiles for men and women; the same conditions underlying the previous figure apply. It is clear that once the parity point is passed after graduation the salary disparity widens through 35 years of experience. The catch-up process begins at 44 years of experience (75 years of age). Catch-up is complete, with the salary differential totally eliminated, at an experience level of 87 (118 years of age). This somewhat amusing mental exercise suggests that experience which does not develop the specific skills measured here provides females with little hope for salary equality. A corollary is that women are more likely to attain parity if their vitae are longer.[15] More of this shortly.

Salary differentials between sexes are examined further in Table 4. There the coefficients from Tables 1 and 3 are manipulated to yield salaries for males and females who possess like skills. Underlying this experiment are givens identical to those for Figure 2 with an added stipulation of 13 years experience. The resulting salary figures may fall short of today's standard because they have not been adjusted for the inflation which followed 1972-73. For all skill combinations listed, females earn less than males. But in contrast to many of the earlier studies in this area, our results suggest substantial differences in the salary relative, depending on the skills possessed. On the assumption that individuals with the same skill package are similar, our results suggest substantial discrimination between male and female faculty currently in administration (0.81) and relatively less for those who have published 3-4 articles (0.91). As can be observed, no clear pattern emerges. In percentage terms, an unskilled female does better relative to her male counterpart than a female administrator with 11-20 articles does relative to hers.

[15]It is not clear what would happen if more women began to publish, however. If a separate labor market exists for them, then an increase in the supply of articles, other things equal, would reduce the return to this activity.

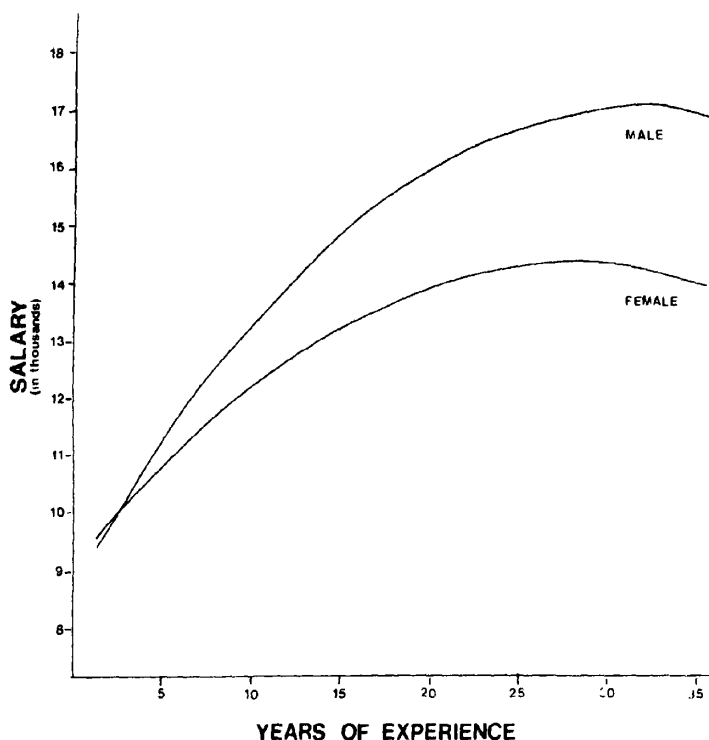



FIGURE 2. EXPERIENCE-EARNINGS PROFILES FOR COMPARABLE MALE AND FEMALE SOCIAL SCIENTISTS

TABLE 4—SALARY COMPARISONS FOR MALE AND FEMALE SOCIAL SCIENTISTS HAVING LIKE SKILLS

Skill Package	Female Salary	Male Salary	Salary Relative[a]
No Skills	$12,923	$14,391	.898
3-4 Articles	13,651	14,953	.913
11-20 Articles	14,091	16,149	.873
Outstanding Teaching	13,341	14,667	.910
Public Service	13,645	15,034	.908
Current Administration	14,453	17,794	.812
11-20 Articles, 5-10 Books	16,135	18,864	.864
11-20 Articles, Current Administration	15,621	19,552	.799
21-50 Articles, 3-4 Books, Current Administration	18,210	21,839	.834
3-4 Articles, Outstanding Teaching Public Service	14,791	15,872	.932[b]

[a]Female salary/male salary.
[b]Minimum salary difference from all possible skill packages.

In absolute terms, however, the skilled female earns a salary greater than that of her unskilled sister. Such findings intimate that studies like those of Hoffman and Johnson and Stafford tend to hide the importance of skill differences in establishing both the existence and the nature of salary discrimination.[16] (It follows that efforts to eliminate discrimination based on equalization of salaries by education and experience levels may create new inequities and inefficiencies.)

Women earn less than comparable men. How many extra credentials are necessary for women faculty to reach salary parity? One way to handle this question is to bestow various combinations of skills on a

[16]The magnitude of the salary differential attributable to discrimination is markedly lower (9 to 20 percent) than those of Hoffman and Johnson and Stafford. While the latter ascribe approximately 40 percent to discriminatory practices, Hoffman assigns 56 to 68 percent of the differential to discrimination.

TABLE 5—SCHEDULE OF EXTRA CREDENTIALS FOR FEMALE SOCIAL SCIENTISTS

Number of Articles	Total Return ($)	*FEAR* Ratio
5	784	.952
6	821	.955
7	858	.958
8	895	.960
9	932	.963
10	968	.965
11	1,004	.968
12	1,041	.970
13	1,077	.973
14	1,113	975
15	1,150	.978
16	1,200	.981
17	1,263	.986
18	1,326	990
19	1,390	995
20	1,453	999
21	1,516	1.003
22	1,579	1.008
23	1,643	1 012
24	1,706	1.017
25	1,769	1.021

female and to compare her consequent salary with that for a male having given skills. The number of possible combinations, however, is quite large, and because of the dichotomous nature of the skill variables exact parity would be difficult to construct. An alternative tack is to derive from the discrete variables a continuous skill measure and to use that measure in locating parity. This is done in Table 5, which is premised on the same conditions as Table 4 The total return column presents the total extra salary resulting from article publication. These figures were computed by linearly interpolating the article coefficients in Table 3 over the midpoints of the publication intervals. The *FEAR* (FEmale ExtrA cRedential) ratio is simply the female's salary inclusive of the return to articles divided by the salary of an *unskilled* male, the latter being $14,391. Thus a female social scientist who publishes 8 articles and has no other credits earns $12,923 + $895 = $13,770, which is 96 percent of an unskilled male's salary. Parity is achieved with 20 articles; put in these terms, the salary disparity seems appreciable.

The final regression result considered is the coefficient for black female faculty: it is large and significant. The repetition of a positive coefficient in Tables 1 and 3 encourages at least a tentative conclusion about the effect of race on salary.[17] From the No Skills salaries in Table 4, it is easy to show that black males and black females earn more than white males, who in turn earn more than white females. White women, not blacks, seem to be the disadvantaged.

V. Concluding Comments

The structure of faculty salaries differs by discipline and sex. These differences are fundamental and cannot be captured by the mere insertion of intercept dummies into an estimating equation. This finding, based on a data base which encompasses a large segment of the academic community, raises doubt about the appropriateness of the fundamental assumptions underlying much of the earlier work on faculty salaries. Of the skills examined here, teaching and public service yield low compensation; publishing and administration carry much larger returns.

Women earn less than men with like characteristics, and this disparity supports Reagan's "dual labor market" hypothesis and extends it to the market for specific skills. Strict inferences about sex discrimination drawn from this conclusion, however, must be accompanied by a key proviso of equal or smaller female supplies. The limited evidence on this point contained in the *ACE* data file is inconclusive. One is nevertheless tempted to ask why separate markets exist for men and women if not because of discrimination.

The salary catch-up phenomenon of Johnson and Stafford is detected by the regressions, but the opportunity provided for women to achieve salary equality is vacuous in the absence of major medical developments to forestall aging. The role of experience in the present skills context differs from that assigned by Johnson and Staf-

[17]The small percentage of blacks in the data (<1 percent) cautions that further evidence is needed before a firm judgment can be made.

ford. In their study experience essentially subsumes the skill variables broken out here for special consideration. With the rewards to select skills articulated, and subject to the constraints imposed by our rather crude measures, women can assume a more active role in the pursuit of salary parity than Johnson and Stafford concede.

Finally the skill differentials identified and analyzed in this paper have important implications for the recent efforts of state legislatures to impose twelve-hour laws and other "accountability measures" on faculty. Legislative fiat will likely be less than totally successful as long as it fails to recognize the nature of the reward structure faculty face. At best such actions will encourage begrudging acceptance; more realistically they will encourage affirmatively evasive action. To the extent that faculty are sensitive to the returns to individual skills, legislative action should be more inclined to success if implemented through the beguilements of the market rather than through coercion.

REFERENCES

A. E. Bayer, "Teaching Faculty in Academe: 1972-73," Office of Research, American Council on Education, Vol. 8, no. 2, 1973.

David G. Brown, *The Market for College Teachers*, Chapel Hill 1965.

———, *The Mobile Professor*, Washington 1967.

E. P. Hoffman, "Faculty Salaries: Is There Discrimination by Sex, Race, and Discipline? Additional Evidence," *Amer. Econ. Rev.*, Mar. 1976, *66*, 196–98.

G. E. Johnson and F. P. Stafford, "The Earnings and Promotion of Women Faculty," *Amer. Econ. Rev.*, Dec. 1974, *64*, 888 903.

D. A. Katz, "Faculty Salaries, Promotions, and Productivity at a Large University," *Amer. Econ. Rev.*, June 1973, *63*, 469–77.

B. B. Reagan, "Two Supply Curves for Economists? Implications of Mobility and Career Attachment of Women," *Amer Econ. Rev. Proc.*, May 1975, *65*, 100–07.

Kenneth D. Roose and Charles J. Anderson, *A Rating of Graduate Programs*, Washington 1970.

Howard P. Tuckman, *Publication, Teaching, and the Reward Structure in Academe*, Lexington 1976.

Pierre van den Berghe, *Academic Gamesmanship*, New York 1970.

Health, Family Structure, and Labor Supply

By DONALD O. PARSONS*

The growing likelihood of major health ifficulties is a principal feature of the ging process, and coping with poor health ıay be the single problem which most disinguishes the economic circumstance of ılder workers from that of the young. leyond the direct utility loss of diminished ıealth, the person in poor health suffers ıther adverse economic and social conse|uences, including a reduction in income as e or she is forced to reallocate time from narket work to health maintenance activiies.[1] A number of economists (Michael Grossman and Lee Benham, Monroe lerkowitz and William G. Johnson, Richard ìcheffler and George Iden, Karen Schwartz, nd Harold S. Luft, 1974, 1975) have ecently attempted to quantify the effect of ıealth status on male labor supply and have ound, not surprisingly, that the effects are arge indeed.[2]

The magnitude of the economic loss for a ;iven severity of illness or accident depends, ɔf course, on the institutional structure within which the individual lives and works. Since the family has traditionally been an important, if informal, health production organization and source of income insurance, the role of the family in conditioning the relationship between health, labor supply, and earnings will be explored in this paper.[3] Obviously the interaction of family structure, health status, and labor supply is quite complex, and I focus below on a few aspects of this issue. They relate largely to the ability of older families to buffer the economic losses imposed by the adult male's poor health.

The main labor supply questions to be considered are the effect of family structure on the male's labor force withdrawal for a given health loss and, for married men, the corresponding sensitivity of the spouse's market activity to the male's health condition. The relative size of the afflicted individual's (and family's) income loss will obviously depend critically on the size of these labor supply responses.[4] The family structure of the male will be characterized simply by the presence or absence of a spouse and the education level of the spouse, if present. Spouse's education is suggested by several health studies which found it to be an important determinant of male health.[5]

Below I consider the health, family structure, and labor supply interrelationships at both theoretical and empirical levels. The paper is organized in the following way. In Section I an informal discussion of the choice structure of families in the face of

*Associate professor of economics and research associate, Center for Human Resource Research, Ohio State University. This research was supported by grants from the National Institute of Child Health and Human Development, National Institutes of Health, Department of Health, Education, and Welfare, and the Manpower Administration, Department of Labor. The study was completed during my tenure as research fellow at the National Bureau of Economic Research. The comments of Michael Grossman and members of the labor workshops at Stanford University and Ohio State University are gratefully acknowledged. David Rahrig provided excellent research assistance.

[1]Some portion of this loss is no doubt insurable although the problems of moral hazard are likely to be quite large in this situation for all but the most obvious physical difficulties.

[2]The study by James N. Morgan et al. (1962) also remains interesting and useful. Other economists have examined the determinants of work days lost due to sickness. See Joseph P. Newhouse, Morris Silver, and Grossman (1972). Since this measure is limited to employed persons it is not a very useful measure for severe health difficulties.

[3]For one interesting effort to consider this problem from a largely sociological viewpoint see Z. Saad Nagi and Linda Hadley.

[4]In this paper I explore only the labor supply consequences of poor health and not wage rate effects. The latter is sufficiently complex to require separate treatment.

[5]Grossman (1976), for example, has found in one sample that the health of an older male is more closely correlated with the education of the wife than his own education.

poor health is outlined. In Section II data from the older male portion of the National Longitudinal Surveys (*NLS*) are used to estimate labor supply functions for married and single men with special attention to differences in poor health responses. A simultaneous model of male labor supply and other family income (chiefly transfer income and the earnings of the wife) is then estimated to determine whether variations in the work hours of males, largely due to health differences, induce any substantial changes in income producing activities by other family members. Finally, in Section III the detailed time budget data on both males and females from the Productive Americans Survey (*PAS*) are used to estimate more precisely the effect of health on total family time allocations. These data provide estimates of the impact of poor health on home production time as well as market time for both husband and wife.

I. The Structure of Family Response to Health Problems

The onset of poor health in the adult male generates a wide range of problems for the individual and the family, and corresponding adjustments in economic behavior. In an appendix to this paper (available upon request) I formally develop a family time allocation model for the adult male in poor health. In this section I briefly summarize some of the important theoretical conclusions (and ambiguities) which arise in that model as a framework for the empirical work which follows.

The adult male's own work time will surely be lower following a decline in health, although the size of the decline is conditioned by a variety of factors. For some illnesses and accidents, market work may be impossible. For many health conditions, however, market work is at least feasible and work hours therefore subject to choice. In these circumstances, the individual and family confront the difficult decision of allocating the male's time between health investments, such as rest, and current income production.[6]

The characteristics of the family are potentially strong conditioning factors for the male's work hour decision. The presence of a wife with the frequently accompanying long-run specialization of market and home tasks may increase the male's responsibilities, inducing him to work in spite of his health difficulties. Most evidence indicates, however, that married men are healthier and recover (to the extent possible) more quickly from negative health shocks, which does not support the notion that they are sacrificing health maintenance for current income.[7] To the extent the wife is able to substitute nursing and personal care activities for the husband's own time in health maintenance, the male in poor health with spouse present should be able to spend more time in the market without suffering adverse health consequences.

The expected reallocation of the wife's time between home and market activity is ambiguous. The onset of poor health in the husband increases the demands for the wife's nursing and personal care services, but also increases the family's demand for income earning activities since few workers are fully compensated for health related work loss. The wife's time allocation choice in this circumstance will presumably depend upon the relative wage rates of husband and wife, the substitution possibilities between the time of husband and wife in his health maintenance, and the entry costs of

[6]A number of researchers including Richard Auster et al., Grossman (1972), and Victor Fuchs have stressed the importance of social phenomena consumption habits, personal relationships, etc. as determinants of health status, in contrast to narrowly defined medical care.

[7]An alternative explanation of this relationship is possible, namely that healthier males can attract and marry higher quality (educated) females. Since these results are for older males, ages 45-59, the simultaneity problem is somewhat diminished. The bulk of these individuals are likely to have developed health problems in the twenty to thirty years since the usual marriage age. See Benham and Finis Welch for a discussion of this issue.

ıe wife into the labor market if she is not ırrently in the market.

The education level of the wife, if present, so has a theoretically uncertain direction f effect on the female's adjustment. More ighly educated women face higher market ages and therefore have more of an incen-ve to substitute time in the market. Alter-atively the efficiency of the wife in pro-iding health maintenance services for the usband should also rise with education. he effect of wife's education on time re-llocation following illness in the husband ltimately rests on the empirical question of hether market or home efficiency rises ıore rapidly with education.

The ambiguities in the theoretical pre-ictions make apparent the need for empiri-al estimates. While the market time of the nhealthy male is likely to fall less when a pouse is present, the change in market time f the wife is uncertain in direction. The ef-ect of the wife's education level is similarly mbiguous a priori. In the next two sections report on empirical efforts to resolve these ssues.

II. The Labor Supply of Older Males: Empirical Results.

A. *Annual Work Hours of Older Males*

In this section, data from the older male ohort (ages 45–59) of the National Longi-udinal Survey (*NLS*) will be explored.[8] The abor supply behavior of the men and, in particular, their differential response of annual work hours to poor health will be estimated.[9] The sensitivity of other income, specifically earned income of other family members, to the work hours of the males is then estimated in a simultaneous frame-work.

In Table 1, column (1) the coefficients of a fairly standard annual work hours model are reported with health status variables in-cluded. The health condition of the in-dividual is characterized by a set of dum-mies *HG*, *HF*, and *HP* which equal one if self-reported health status is judged to be good, fair, or poor, respectively; zero other-wise.[10] The reference health status is excel-lent health, *HE*. The health variables, par-ticularly the presence of poor health, strongly influence labor supply. Poor health status, for example, implies a reduction of 1300 hours or 65 percent of a standard work year of 2000 hours.

The nonhealth coefficients are of inde-pendent interest. Higher skilled workers generally work longer hours as do married men, particularly married men with highly educated wives. A married man with a wife who attained twelve years of schooling can be expected to work about 260 hours more per year than a single man, health and other factors constant. Each dependent adds another 20 hours.

To answer the question of whether mar-ried men have a different labor supply re-action to poor health than do single men, the total sample was separated into mar-ried and single subsamples and the regres-sion model with health variables reesti-mated. The coefficient estimates are reported in Table 1, columns (2) and (3), for the married and single subsamples, respec-tively.

The differences in behavior by health status are dramatic with hours reductions (relative to a base of excellent health) of

[8]The *NLS* is a national longitudinal survey of the labor market characteristics of four age cohorts of about 5000 persons each. The survey is representative of the *U.S.* population except for a substantial over-sampling of blacks (30 percent of total). See Herbert S. Parnes et al.

[9]Annual hours were chosen because of interest in measuring roughly an earnings effect of health. Other researchers have estimated health effects on labor force participation, weeks worked per year, and hours worked per week. Both Joseph Davis and Richard Scheffler and George Iden report that weeks worked are more sensitive to poor health than are hours per week Paul Burgess and Jerry Kingston report mod-estly greater duration of unemployment for workers in poor health.

[10]A classic piece by Nagi compares self-reported health condition with a doctor's opinion and finds a substantial correspondence of the two. Deviations where systematic are not necessarily in the intuitively predictable direction.

TABLE 1—ANNUAL MARKET HOURS OF MEN, 45–59, IN 1966, TOTAL AND BY MARITAL STATUS[a]

Variable	(1)	(2)	(3)
Skill			
S^b	23.04	27.88	−11.35
	(2.15)	(2.40)	(0.38)
S^2	−1.42	−1.67	0.48
	(2.60)	(2.88)	(0.28)
Blue[c]	−330.68	−338.20	−277.26
	(14.13)	(13.76)	(3.73)
Family			
M^d	166.29		–
	(3.00)		
$M \times SW^e$	8.02	8.62	
	(1.76)	(1.87)	
Depend[f]	21.73	18.51	53.85
	(3.32)	(2.79)	(1.82)
Demographic			
Black[g]	−169.33	−174.83	−136.70
	(6.30)	(6.10)	(1.72)
Age[h]	−9.02	−10.19	−2.21
	(3.35)	(3.60)	(0.26)
Market			
Unemp[i]	−11.83	−10.13	−21.50
	(2.39)	(1.87)	(1.67)
Health			
HE^j	–		
HG^k	−48.52	−34.43	−157.75
	(1.93)	(1.32)	(1.82)
HF^l	−203.30	−172.09	−391.21
	(6.26)	(5.03)	(3.87)
HP^m	−1301.9	−1222.2	−1677.0
	(28.01)	(24.45)	(12.83)
Constant	2658.8	2846.4	2528.3
	(16.41)	(16.38)	(5.07)
R^2	0.25	0.23	0.30
Sample	Total	Married, Spouse Present	Other than Married
Sample Size	4444	3865	569

Source: National Longitudinal Survey.

[a] Absolute value of *t*-ratios are in parentheses
[b] Years of schooling of man
[c] Dummy equal to one if current or last occupation blue collar.
[d] Dummy equal to one if married, spouse present
[e] Years of schooling of wife, if present.
[f] Number of dependents, excluding wife if married.
[g] Dummy equal to one if race nonwhite
[h] Age of husband.
[i] Local labor market unemployment rate.
[j] Dummy equal to one if self-reported health status excellent.
[k] Dummy equal to one if health good.
[l] Dummy equal to one if health fair.
[m] Dummy equal to one if health poor.

34, 172, and 1222 hours for good, fair, and poor health, respectively, for married men and 158, 391, and 1677 hours for men without spouse present. The differential work hour reduction of single men then is 124, 219, and 455 annual hours for the three lesser health categories. An F test of the hypothesis that the two regression structures have equal coefficients can be rejected at the 1 percent level.[11] In terms of a fraction of a normal work year (2000 hours), married men in poor health are forced to contract their annual hours in the market by 61 percent while single men contract their work hours by 84 percent. Single men can then expect to suffer a reduction beyond that of married men of an additional 23 percent of their original earnings when poor health strikes. This suggests that the wife's nursing and care services are a substantial substitute for the husband's time in the provision of health maintenance services. As a result, the married man can remain more in the market for given health levels.

The interesting question arises whether the "quality" of the wife measured by her level of education will also affect the degree of labor force withdrawal of unhealthy married men. Theoretically the answer is ambiguous since schooling is likely to increase the female's efficiency in health maintenance, implying a reduction in male withdrawal from the market, and also likely to increase her market wage, with the opposite effect on male labor supply. Regressions identical to that reported in Table 1, column (2), were run separately for married men whose spouses had twelve or more years of education and those whose spouses has less than twelve years of education. The results, not reported here, suggest very little difference in male labor supply reduction as men with highly educated wives withdrew 0, 152, and 1169 hours when health was good, fair, and poor; while men with low educated wives withdrew 69, 191, and 1240 hours per year. The maximum difference is

[11] A test of the equality of the health variables alone was also undertaken and rejected at the 1 percent level.

never greater than 71 hours per year or less than two normal weeks.

B. *A Simultaneous Model of Male Work Hours and Other Income*

The empirical estimates of the preceding section indicate that the labor supply (and therefore earnings) of older males drops sharply with poor health. An important social question remains, however, of how adequately transfer income and earned income of other family members compensate for this loss. In this subsection I attempt to provide estimates of the effects of other income on male labor supply, and effects of male labor supply on other income, using simultaneous methods. This should allow us to determine whether the other income of the family conditions labor supply behavior of the male and, more importantly, whether male labor supply reductions induce increased transfer income and earnings by the spouse.

In the simultaneous model below, total other family income was divided into two components: the sum of wealth income and nonwork related transfer payments; and the sum of work related transfer income and earned income of other family members. The former was assumed to be exogenous to the system and was initially treated as an explanatory variable in labor supply, although it was later dropped because it showed positive wealth effects on labor supply.[12] The second sum is assumed to be endogenous in this system and will be called simply "other income."

The content of the vectors of exogenous variables which would be expected to enter the male work hour and other family income equations remains to be specified. The exogenous variables in the work hours equation have been discussed at length above; the relationship will be assumed similar to that reported in Table 1. The exogenous variables in the other income equation presumably include factors which influence the earning power of the wife (schooling, her health, etc.) as well as factors likely to affect the size of transfer flows (number of dependents and urban residence).

Turning to the estimation results, one will find the reduced form estimates of hours and other income reported in Table 2, columns (1) and (2). The estimates of coefficients in the work hours regression are not much different from the earlier results reported in Table 1, although the coefficient on one additional variable, wife's health for married men is interesting. If the wife has an activity limiting health problem, the husband works an average of about 100 hours less than he otherwise would, presumably because he must increase his home production time. We will consider this issue at greater length in the next section.

The other income reduced form estimates in column (2) have not been discussed above and bear closer examination. Other income is substantially higher for men with healthy, well-educated spouses. The significant quadratic form for wife's schooling in this equation suggests that other income increases at an increasing rate with wife's education. If the wife has an activity limiting health problem, other family income is reduced by an average of more than $500. Other income, however, rises with various indices of poor health in the husband, presumably due to some combination of welfare and adjustments in other family earnings. The reasonably small sizes of the health effects on transfers and other family earnings, under $1,000 in the case of poor health in the husband, suggest that the effect of male labor supply on other income is substantial but well short of fully compensating the workers in poor health. Urban residence corresponds to about $300 more in other income for the family.

Two-stage least square estimates of the structural equations are reported in Table 2, columns (3) and (4). The coefficient of male hours on other family income is significant and indicates that other income increases by about $0.75 for every one-hour reduction in male hours worked. The coefficient of other income in the hours equation, however, is insignificant and indeed posi-

[12] A complete model would no doubt treat assets and asset income as an endogenous variable.

TABLE 2—ANNUAL WORK HOURS AND OTHER INCOME OF MALES, 45-59, IN 1966, REDUCED FORM AND STRUCTURAL EQUATIONS[a]

	Reduced Form		Structural[b]	
Dependent Variables	Hours[c] (1)	Other Income[d] (2)	Hours (3)	Other Income (4)
Hours*[c]		-		−0.74
				(6 38)
Other Income*[d]	-	-	0.03	-
			(1 33)	
Skill				
S	38.28	144 3	44.17	3.75
	(3 15)	(3 72)	(3.68)	(0.29)
S^2	−1 74	−7 70	−2.08	-
	(2.70)	(3 76)	(3.37)	
Blue	−300.1	344.9	−335.3	-
	(11 35)	(4.09)	(12.23)	
Family				
M	130 4	1766	232 2	1532
	(1.24)	(5.27)	(4 74)	(4.63)
$M \times SW$	27 39	−238.9		−126.4
	(1.37)	(3.74)		(2.04)
$M \times SW^2$	−0 96	21.21	-	15.40
	(0 96)	(6.64)		(5.02)
$M \times HW^e$	−123.5	−551.0	-	−126.4
	(3.94)	(5.50)		(2.04)
Depend	13.64	−85.08	18 10	−87.97
	(1.91)	(3 74)	(2.43)	(3.97)
Demographic				
Age	−8 75	10 81	−10.41	-
	(2 91)	(1 13)	(3 41)	
Urban[f]	155 0	281 7		159 3
	(5.93)	(3.38)		(1.88)
Market				
Unemp	−18 21	27 29	−17 86	−44 98
	(3 24)	(1 52)	(3 10)	(2 45)
Health				
HE	-	-		-
HG	−62.40	183 3	−57 94	
	(2 22)	(2.04)	(2 02)	
HF	−221.4	342.4	−220.3	-
	(6 08)	(2 94)	(5.95)	
HP	−1303	999.0	−1315.0	-
	(24.63)	(5.92)	(23.30)	
Constant	2596	−1213	2558	1776
	(14.56)	(2.13)	(14.06)	(6.74)
R^2	0 26	0.12	-	-

Source: National Longitudinal Survey.

[a]Absolute values of t-ratios are in parentheses. Notation not defined here can be found in Table 1. The sample size is 3,428

[b]The structural equations were estimated using two stage least squares procedures.

[c]Annual work hours. Asterisk denotes instrument derived from column (1).

[d]Total other family income less asset and nonwork related transfer income. Asterisk denotes instrument derived from column (2).

[e]Dummy equal to one if wife has an activity limiting health condition.

[f]Dummy equal to one if urban residence.

ive. Taken at face value these results suggest that male hours affect the earnings behavior of other family members and the flow of transfer payments, but are not themselves affected by the size of these other income flows. The latter result, however, must be held with some caution since the instrument used for other income has an R^2 of only 0.12, suggesting that almost 90 percent of the variation in other income has been discarded in the estimation process. The instrument may not be very useful in this case. Most of the other coefficients are not substantially changed from the reduced form estimates.

To throw some light on the question of whether it is other earned income or transfer payments that primarily account for the increase in other family income when male work hours are reduced, the simultaneous estimation was repeated for married men only, and other income limited to other *earned* family income. The reduced form equation for other earned income in this model (not reported here) indicates that this income is fairly insensitive to the health status of the male. Income increases by $130, $235, and $182 as health drops from excellent to good, fair, and poor, respectively. In the structural equations, the estimated coefficient of hours on other family income is sharply reduced in magnitude from the preceding model. A loss of one hour worked by the male is offset by only about $0.23 increase in the earned income of other family members. Apparently most of the income compensation for health induced variation in husband work hours results from transfer payments and not from work adjustments of other family members.

Dividing the sample according to wife's education ($SW \geq 12$, $SW < 12$) and reestimating the equations does not materially alter these conclusions (the full results are not reported here). The insurance value of more highly educated wives is substantially, if not statistically significant, above that of less educated wives. For every reduction of one hour in male work hours, other family earned income rises by $0.49 for highly educated wives and only $0.14 for less educated wives with respective t-values of 1.72 and 1.06. While interesting, in neither case is the effect statistically significant at customary levels.

Apparently other family members do not go into the market in a strong and systematic way when the husband falls ill. This may partially result from the fixed costs of either entering the market if one does not currently have a job or of increasing one's hours if one is already in the market. The model of the preceding section, however, suggests this result might be due to the increased home time demands of the wife while the husband is ill. The question of intrafamily time allocation is pursued in more detail in the next section with a data set which allows exploration of the behavior of home production hours as well as market hours.

III. Health and Intrafamily Time Allocations

In this section I examine the impact of health problems on the nonmarket as well as market time of husbands and wives using data from the Productive Americans Survey (*PAS*).[13] This survey includes time budget data on productive work hours in the home (cooking, cleaning, house maintenance) as well as in the market. The health measures are less complete in this survey. Only information on the existence of a work limiting health condition is available, a variable that must be viewed with some caution in a labor supply study.[14] Nonetheless, the results should provide some useful insight into the complete family time allocation when husband or wife fall ill or are disabled.

[13]The *PAS* is a single survey carried out in early 1965 with observations on 2214 families of all ages. See Morgan et al (1966).

[14]The correspondence between a measure of this sort and the one on health status used earlier is quite rough. For the *NLS* sample which contains both measures, almost all respondents (95.9 percent) who reported themselves in poor health also reported a work limiting health condition. Many with a work limiting condition, however, felt themselves in good and even excellent health.

TABLE 3—HEALTH COEFFICIENTS IN ANNUAL PRODUCTIVE HOURS REGRESSIONS: MARRIED MEN, 45–64, AND MARRIED WOMEN, 40–64[a]

Variable	Market Work (1)	Home Work (2)	Total Work (3)
	Married Men		
HLA-H[b]	-695.24	2.72	-692.51
	(8.91)	(0.07)	(8.58)
HLA-W[b]	31.96	193.68	225.64
	(0.28)	(3.55)	(1.93)
	Married Women		
HLA-H	136.56	-28.31	108.25
	(1.63)	(0.30)	(1.10)
HLA-W	-361.14	-20.85	-381.99
	(2.89)	(0.15)	(2.59)

Source: Productive Americans Survey.

[a]Absolute values of *t*-ratios are in parentheses. Sample size was 605 for male sample, 737 for females. Other variables included in the regression are age, schooling, blue-collar status, and race.

[b]Dummy equal to one if health limits or prevents work, zero otherwise. Hyphen *H* denotes a husband variable, hyphen *W* a wife variable.

In Table 3 I report the estimated annual market and home work hours effects of the presence of an activity limiting health condition for the husband (*HLA-H*) and the wife (*HLA-W*). In column (1), estimates of market work are reported. Consistent with the results reported in the preceding section, the husband's poor health forces a reduction in his work hours of about 700 hours annually. His wife's health problems induce a modest and statistically insignificant increase in his work hours of about 30 hours annually. This is inconsistent with the significant and negative coefficient of wife's health on market hours in the *NLS* sample (see Table 3, column (1)) of about 115 hours. The impact of health on the male's home work (column (2)) corresponds with expectations. The husband's health status has no effect on his home work hours although poor health in his wife leads to a statistically significant increase in his home work hours. On average the husband seems to increase his home work by about 200 hours when his wife develops a health problem.

Comparable regression results are reported for married women, 40–64, in Table 3. For married women, poor health involves a reduction in market hours of about 361 hours annually and no significant change in home work hours. Total work hours, both market and home, then drop by about 380 hours. The poor health of the husband leads to an increase of about 140 hours in market work and a small and insignificant decrease in home hours. The latter is made possible, one might conjecture, by the ability to substitute tasks within home hours—for example, nursing for cleaning.

The combined effect of poor health of the husband on total family work hours is a loss of 560 market hours and unchanged home hours. The effect on family hours of poor health of the wife is a loss of about 330 market hours and a gain of 173 home hours. The husband works about 600 hours less in total when he is ill, the wife about 400 hours less when she is ill. When the spouse is ill, men increase their work hours and decrease their leisure by about 200 hours, largely in home production, while women increase their work hours by about 100 hours, largely in market work.

Separating the two samples by the schooling of the female (less than twelve years and greater than or equal to twelve years) revealed only modest differences by type of household. Wife's health difficulties did induce a somewhat different time allocation by the male in these two groups. Males with less educated wives tended to increase home work hours by 279 hours while reducing market hours by 91 hours. Males with more educated wives increased home hours by only 73 and *increased* market hours by 207. Only the result for home time of the husbands of less educated women was statistically significant however. Apparently males in low education families find it more economical to substitute their own time in household activities when their wives fall ill, while males in higher education families substitute market goods.

Similarly female market work is more sensitive to health problems in the spouse among well-educated women. The positive market work hour effect of husband's poor health, noted in the sample as a whole, is solely due to the response of highly educated wives who increase their work in the market by 270 hours. The less well-educated wives did not increase their market hours at all when the husband encounters a health difficulty. This differential response is perhaps due to the easier access of well educated females to the labor market.

IV. Conclusion

The main objective of this study has been the exploration of the interrelation of health and allocation of time within the family. Particular attention is paid to health effects on the joint labor supply of husbands and wives and to the differential labor supply responses to poor health of married and single men. The impact of health on home production hours of the family is also considered. The results are of more than academic interest as they give an indication of how well older individuals and families can economically cope with poor health.

The empirical analysis of the labor supply of older men (ages 45–59) from the *NLS* demonstrates the importance of family structure in conditioning the labor supply response to poor health. An individual in poor health works on average 1300 hours less per year than similarly educated (and aged) men in excellent health. The work hours reduction for married men is however 450 hours less annually than for single men (or more precisely men with no spouse present). Considered from a standard 2000 hour year the decline in annual hours for single men in poor health is 84 percent of a full-employment year while only 61 percent for married men. This evidence is consistent with the notion that married men can marshal resources other than their own time, particularly their wives' time, to augment the health of the male. This effect is largely independent of the wife's education level.

A simultaneous model of male labor supply and other family income is also estimated on the *NLS* data to determine the effect of other family income on male labor supply and male labor supply on other family income. The models suggest that other income does not have a substantial effect on labor supply but that male labor supply has a significant effect on other family income. Total other income is estimated to increase about $0.75 for each one hour reduction in work hours of the older male. About two-thirds of this subsidy comes from social welfare sources, one-third from increased earnings of other family members. Other family income increases substantially only in households where the wife has high levels of education.

Finally, in Section III time budget data from the *PAS* are used to trace out more fully health effects of family time decisions. For both husbands and wives, one's own health problems appear to lead to substantial market time withdrawals (about 700 hours and 350 hours, respectively) while home work hours remain unchanged. As one might expect, illness in one's spouse leads to quite different time allocation responses as men increase their home production time, women their market work time. These work time increases appear to come largely from leisure time in both cases.

Clearly the results reported in this paper represent only the beginning of the investigation of the family as an informal health service organization. A particularly crucial element that requires further study is the role of wage rate variations and other productivity effects of poor health in this choice structure.[15] A number of other researchers have found significant racial differences in labor supply response to poor health so this again may be a useful area for further exploration.[16]

[15] Again see Grossman and Benham.

[16] See Berkowitz and Johnson, Scheffler and Iden, and Luft (1975)

REFERENCES

R. Anderson and L. Benham, "Factors Affecting the Relationship between Family Income and Medical Care Consumption," in Herbert E. Klarman, ed., *Empirical Studies in Health Economics*, Baltimore 1970.

R. D. Auster et al., "The Production of Health: An Exploratory Study," *J. Hum. Resources*, Fall 1969, *4*, 411-36.

G. Becker, "A Theory of the Allocation of Time," *Econ. J.*, Sept. 1965, *75*, 493-517.

L. Benham, "Benefits of Women's Education within Marriage," *J. Polit. Econ.*, Mar./Apr. 1974, Part II, *82*, S57-S71.

M. Berkowitz and W. G. Johnson, "Health and Labor Force Participation," *J. Hum. Resources*, Winter 1974, *9*, 117-28.

P. L. Burgess and J. L. Kingston, "The Effect of Health on Duration of Unemployment," *Mon. Lab. Rev.*, Apr. 1974, *97*, 53-54.

J. M. Davis, "Impact of Health on Earnings and Labor Market Activity," *Mon. Lab. Rev.*, Oct. 1972, *95*, 46-48.

Victor R. Fuchs, *Who Shall Live*, New York 1974.

Michael Grossman, *The Demands for Health: A Theoretical and Empirical Investigation*, New York 1972.

——, "The Correlation Between Health and Schooling," in Nestor E. Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, vol. 40, New York 1976.

—— **and L. Benham,** "Health, Hours, and Wages," in Mark Perlman, ed., *The Economics of Health and Medical Care*, London 1974.

H. S. Luft, "Education, Activity Limitations, and Labor Force Participation: A Source of Bias in Human Capital Analysis," mimeo., Stanford Univ. 1974.

——, "The Impact of Poor Health on Earnings," *Rev. Econ. Statist.*, Feb. 1975, *57*, 43-57.

James N. Morgan et al., *Productive Americans: A Study of How Individuals Contribute to Economic Progress*, Ann Arbor 1966.

——, *Income and Welfare in the United States*, New York 1962.

Z. S. Nagi, " Congruency in Medical and Self-Assessment of Disability," *Indust. Med. and Sur.*, Mar. 1969, *38*, 27-36.

—— **and L. W. Hadley,** "Disability Behavior: Income Change and Motivation to Work," *Indust. Lab. Rel. Rev.*, Jan. 1972, *25*, 223-33.

J. P. Newhouse, "Determinants of Days Lost from Work due to Sickness," in Herbert E. Klarman, ed., *Empirical Studies in Health Economics*, Baltimore 1970.

Herbert S. Parnes et al., *The Pre-Retirement Years: A Longitudinal Study of the Labor Market Experience of Men*, Vol. I, Washington 1970.

R. M. Scheffler and G. Iden, "The Effect of Disability on Labor Supply," *Ind. Lab Relat. Rev.*, Oct. 1974, *28*, 122-32.

K. Schwartz, "Early Labor-Force Withdrawal of Men: Participants and Nonparticipants Aged 58-63," *Soc. Secur Bull.*, Aug. 1974, 24 38.

M. Silver, "An Economic Analysis of Variations in Medical Expenses and Work-Loss Rates," in Herbert E. Klarman, ed., *Empirical Studies in Health Economics*, Baltimore 1970.

F. Welch, "Comment: Benefits of Women's Education within Marriage, by Lee Benham," *J. Polit. Econ.*, Mar./Apr. 1974, Part II, *82*, S72-S75.

"National Longitudinal Survey 1966-69," *(NLS)* Ohio State Univ.

Trade Creation and Trade Diversion in the Council of Mutual Economic Assistance: 1954–70

By JOSEPH PELZMAN*

The Council of Mutual Economic Assistance[1] (*CMEA*) has been in existence since January 1, 1949. Despite the fact that its creation[2] was attributed to the establishment of the Organization for European Economic Cooperation (*OEEC*), its declared ultimate goal was the promotion of a process of integration among the East European countries.[3]

For the Soviet Union, economic integration is primarily a means of increasing its political and economic control over the other *CMEA* member states. On the other hand, for the more developed members of *CMEA*, economic integration is a natural outgrowth of their desire to industrialize and maximize the economic gains from trade and cooperation.[4]

Beginning in the early 1960's, the shift from an extensive to an intensive growth policy in response to the decline in growth rates created a drive towards economic integration within *CMEA*. The desire to increase the static gains from international trade was further prompted by its expected contribution to rapid industrialization and efficiency. This shift to intensive growth combined with the desire for rapid industrialization has also meant greater decentralization of economic decision making and the use of limited market mechanisms.[5]

The *CMEA* as it exists today differs from the *EEC* customs union in one major respect. It does not rely on a clearly defined common external tariff. A proxy of such a tariff, however, originates in the annual bilateral negotiations between the *CMEA* member states. Consequently, it is quite possible to find the existence of trade creation and/or diversion as effects of economic integration within *CMEA*.

The analysis of trade creation and/or diversion in this study is of an *ex post* type. The model utilized to determine these effects is a cross-sectional trade-flow model of the type developed by Jan Tinbergen, Pentti Poyhonen, Kyosti Pulliainen, and Hans Linnemann. Using this cross-sectional trade-flow equation to empirically test the integration effects of *CMEA* we initially pool the cross-sectional and time-series data for both aggregate and disaggregate trade flows. In the case of disaggregate commodities, because we cannot rule out the possibility that the regression disturbances in different equations are mutually correlated, we use the estimating procedure de-

*Assistant professor of economics, University of South Carolina. I am grateful to the managing editor of the *Review* and an anonymous referee for their helpful comments on an earlier draft of this paper. Responsibility for any errors is solely mine.

Substantial portions of this article were previously published in the *ACES Bulletin*, Vol. XVIII, No. 3, Fall 1976, and this material appears here with the permission of the Association for Comparative Economic Studies. Discussion of the article and a reply appear in the same journal, Spring 1977.

[1] The *CMEA* member countries are Bulgaria, Czechoslovakia, East Germany, Hungary, Poland, Romania, and the *USSR*.

[2] The most plausible reasons for *CMEA*'s origin are presented in Michael Kaser, chs. 1, 2, and in I. T. Berend, pp. 15–17.

[3] The communique published on January 22, 1949 declared that the *CMEA* was created "on the basis of equal representation and with the task of exchanging economic experience, technical aid and rendering mutual assistance with respect to raw materials, foodstuffs, machines, equipment, etc. . ." Heinz Kohler (pp. 377–95).

Economic integration is defined as a process aimed at reducing the disparity between scarcities in the various *CMEA* countries by eliminating obstacles to trade.

[4] See the author (1976a, ch. 3).

[5] For a further discussion of the links between international trade, industrialization, and the reforms, see the author (1976a, ch. 2).

veloped by Arnold Zellner (1962, pp. 350–52). We then begin to test our hypothesis that the linear regression system obeys two separate regimes. The use of Quandt's maximum likelihood technique (1958) and likelihood ratio test proves to be a superior statistical procedure than the use of dummy variables in determining the first year in which integration effects occurred.[6]

After the existence of a structural break has been shown, we proceed to reestimate the trade-flow equation for a stable pre-integration period. In order to make a proper projection of the trade creation and/or diversion effects, this equation is recalculated, leaving out the trade preference variable.

I. The Model and Procedure

The commonly defined integration effects are trade creation (*TC*), trade diversion (*TD*), and gross trade creation (*GTC*).[7] The *TC* effects refer to the emergence of new flows of trade among the partner countries replacing domestic production; *TD* refers to the replacement of nonpartner imports (low-cost products) by partner country imports (more costly products). The *TD* and *TC* effects combined result in *GTC*, which signifies a growth in trade among the member countries, regardless of the reason for this growth.

A large number of empirical models dealing with these *ex post* measures of integration are available. One such model focuses on the market shares of imports in apparent consumption (see Edwin Truman, pp. 206–12). Another is based on import demand equations with one single national variable (see Balassa, 1967, pp 5–11). A third model reconstructs the no-integration or normal level of trade based on demand equations using multiple regressions (see Mordechai Kreinin, 1969, pp 274–76).

While statistically and logically sound procedures, none of these approaches could be used to measure the integration effects of *CMEA*. In general, the requirement in terms of the amount and quality of data necessary for all these models could not be met from the data available.[8] To measure the effects of the *CMEA* on the trade flows of its member countries would require the use of a modified gravity trade-flow model, which does not directly incorporate prices.

The trade-flow equation is:

$$(1)\quad \log X_{ij} = g_0 + g_1 \log Y_j^N + g_2 \log Y_i^N + g_3 \log N_j + g_4 \log N_i + g_5 \log D_{ij} + g_6 \log P_{ij} + \log e_{ij}$$

where X_{ij} = the dollar value of i's exports to j

$Y_i^N; Y_j^N$ = the nominal *GNP* of country i and j in *U.S.* dollars

$N_i; N_j$ = the populations of country i and j

D_{ij} = the distance between the commercial centers of the two countries (geographic distance)

P_{ij} = a dummy preference variable reflecting membership in the *CMEA*. The value 2 is assigned to intra-*CMEA* trade while the value 1 is assigned to inter-*CMEA* trade flows

log refers to natural *logs*

This general equilibrium reduced-form model specifies that trade between country i and j is determined by the relative size of their foreign sectors. In turn, country i's potential foreign supply depends on its national product Y, and on the ratio between production for the domestic market and production for foreign demand explained by differences in population. Given economies of scale, the larger N is, the larger will be the domestic market to foreign market ratio, and the smaller the potential export supply of the country. The variables Y_j and N_j together determine the potential import demand for country j for

[6]The use of dummy variables to measure the *EEC* effects was presented by Norman Aitken. The procedure used here is considered superior to that used by Aitken because it not only identifies the shift in structure but also allows one to create a confidence interval around that switching point. Moreover, this statistical procedure associates the structural shifts with our a priori hypotheses of the political events within *CMEA*.

[7]See Bela Balassa (1967, p. 5).

[8]For a discussion of the data limitations and sources see the author (1976a, pp. 99–102, and Appendix A)

a similar argument. The D_{ij} is a proxy variable for natural trade resistance. Consequently, D_{ij} along with N_i and N_j is hypothesized to have a negative effect on X_{ij}. The dummy variable P_{ij} is used to reflect membership in the *CMEA* group. The estimated coefficient on the dummy variable measures to what extent intra-*CMEA* trade flows were augmented.

A. *Aggregate Trade Flows*

The aggregate trade flow sample consists of 350 trade flows per year for the 17-year period, 1954–70. The trade-flow matrix will therefore consist of 5950 observations on 8 variables (1 dependent variable and 7 independent variables including the constant term).[9]

This matrix can be written as a system of 17 equations, where the μth equation can be represented as:

$$(2) \qquad \log X_{ij,\mu} = \log Y_\mu g_\mu + \log e_\mu$$

where $\log X_{ij,\mu}$ is a 350 x 1 vector of observations on the μth dependent variable, Y_μ is a 350 x 7 matrix of observations on 7 independent variables, g_μ is a 7 x 1 vector of regression coefficients and $\log e_\mu$ is a 350 x 1 vector of lognormally distributed error terms, with $E(\log e_\mu) = 0$. The system of which (2) is an equation is:

$$(3) \quad \begin{bmatrix} \log X_{ij,1} \\ \vdots \\ \log X_{ij,T} \end{bmatrix} = \begin{bmatrix} \log Y_1 0 \ldots\ldots 0 \\ \vdots \qquad\qquad \vdots \\ 0 \ldots\ldots \log Y_T \end{bmatrix} \begin{bmatrix} g_1 \\ \vdots \\ g_T \end{bmatrix} + \begin{bmatrix} \log e_1 \\ \vdots \\ \log e_T \end{bmatrix}$$

where T = 17.

Running a pooled regression over all time periods and all X_{ij} we begin to test our hypothesis that the linear regression system obeys two separate regimes.[10] I believe that a major break in the system should have occurred after the signing of the "Basic Principles" in June 1962. Furthermore, there is some information with respect to joint planning in 1958, 1959[11] which leads one to believe that another break may have occurred at that period as well. Each of these breaks is understood to represent structural changes leading to an increased state of integration within *CMEA*.

The actual procedure to test for the location of this unknown breaking point involves the use of Quandt's maximum likelihood technique and likelihood ratio test.[12]

[9]Our sample includes the following *CMEA* countries: Bulgaria, Czechoslovakia, East Germany, Hungary, Poland, Romania, and the *USSR*. The Western countries considered are the following: Austria, Belgium-Luxembourg, Canada, Denmark, Finland, France, West Germany, Greece, Iceland, Ireland, Italy, Japan, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States, and Yugoslavia. There are 308 inter-*CMEA* trade flows and 42 intra-*CMEA* trade flows.

[10]A number of factors compelled me to choose this technique. First, the exclusion of the price variable in this trade-flow model implies that the market clearing quantity depends on demand and supply factors but not on the price variable. Linneman, therefore used data averaged over a 3-year period to reflect this equilibrium characteristic. However, Helen Junz and Rudolf Rhomberg, p. 452, have found that data over a longer time period should be used to eliminate the influence of short-run price changes. Secondly, the cross-section equation alone is static, thus paying no attention to the development of trade over time (Tinbergen, p. 263). However, we are interested in capturing structural shifts which might develop in the long run because of integration.

[11]In December 1958, a session meeting in Prague implemented specialization agreements in the production of chemicals and joint construction of an oil pipeline. Later in 1959, agreements were reached with respect to specialization in rolled products, mining machinery, civil engineering, oil refining bearings, and rolling mill equipment. After the "Basic Principles" in 1962 *CMEA* made a very important decision to endorse greater specialization in production and international trade. See Kaser (pp. 153–74) and the author (1976a, ch. 3) for a further discussion of intra-*CMEA* cooperation.

[12]To find the best estimate of this break t^* we choose the value of t for which $L(t)$ reaches the highest maximum.

$$L(t) = -T \log \sqrt{2\pi} - t^* \log \hat{\sigma}_1 - (T - t^*) \log \hat{\sigma}_2 - T/2$$

where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the standard errors of the estimates of the left-hand and right-hand regressions,

After the existence of a structural break has been shown, we proceed to reestimate the trade-flow equation for a stable period prior to the break (which we attribute to integration). In order to make a proper projection, this equation is recalculated, leaving out the trade preference variable. Projection estimates which are based on this equation are made on the basis of the usual assumption that the effect of changes in competitive position and trade liberalization on trade has been small relative to the effects of integration.

The difference between the actual intra-*CMEA* trade flows and the hypothetical intra-*CMEA* trade flows of *CMEA*'s preintegration structure is taken to be indicative of the *GTC* effects. The difference between the actual inter-*CMEA* trade flows and the preintegration inter-*CMEA* trade flows will indicate the *TD* effects. The resulting difference between the *GTC* and *TD* effects will be indicative of the *TC* effects.

B. *Disaggregate Trade Flows*

The disaggregate trade-flow sample consists of 37 commodity classifications for the years 1958 to 1970. The total number of trade flows per year and commodity will be 330. The trade-flow matrix will therefore consist of 4290 observations on 8 variables per commodity.[13]

As in the case of aggregate trade flows, we expect each of the equations to satisfy the assumptions of the classical normal linear regression model. However, since we are dealing with disaggregate commodities we cannot rule out the possibility that the regression disturbances in different equations are mutually correlated. Specifically, we contend it is possible that there may exist some common factors that affect the trade decision of countries for a specific commodity in a given period.[14] Thus there may exist a link between the mth and the pth equation and that link is represented only in the covariance of the disturbances of the mth and the pth equation. Because this link is so subtle, we can call this system of T equations (3) a system of "seemingly unrelated regression equations."

Our assumption of correlation between equations, therefore, suggests that an efficient estimation of our model of reduced-form equations, where each endogenous variable is a function of a set of exogenous variables and the only link between the mth and pth equation is σ_{mp}, is the procedure developed by Zellner (1962, pp. 350-52). Essentially, this procedure regards (3) as a single equation regression model and applies Aitken's generalized least squares.

The procedure involved in testing for the location of the unknown breaking point and reestimation of the trade-flow equation is identical to that used for aggregate trade

---

respectively. Once we determine the best estimate of t^* we proceed to test the hypothesis that no switch occurred during the period in question Our alternative hypothesis is that one switch occurred specifically at t^* given by the maximum maximorum of $L(t)$. For this purpose we use a likelihood ratio test derived by Quandt. The likelihood ratio λ is defined as:

$$\lambda = \frac{\hat{\sigma}_1^{t^*} \hat{\sigma}_2^{T-t^*}}{\hat{\sigma}^T}$$

where $\hat{\sigma}$ is the standard error of the estimate for a regression taking into account all observations, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the standard errors of the estimates of the left- and right-hand regressions, respectively, and where t^* has been chosen so as to minimize λ. For large T, $-2 \log \lambda$ is asymptotically distributed as χ^2. For a detailed discussion of the test see Quandt (1958, pp. 875-76). A derivation of a generalized maximum likelihood technique may be found in the author (1976b, pp. 28-30).

[13]The intra-*CMEA* disaggregate foreign trade data which we possess is limited to Soviet trade data by three-digit *Uniform CMEA Foreign Trade Commodity Nomenclature* (*CTN*) for 1954-1970, Czech foreign trade data by three-digit *CTN* from 1958 to 1968 and by *Standard International Trade Classification-Revised* (*SITC*) from 1968 to 1970. Our Polish data in three-digit *CTN* detail is, however, limited to the years 1964-67. We are thus faced with the problem that the Soviet Union and Czechoslovakia are the only *CMEA* countries which publish a full set of disaggregate trade flows for the years 1958-70. The solution to this data availability problem is to employ the mirror statistics of Czechoslovakia and the Soviet Union. For a more detailed discussion of the problems involved see the author (1976a, Appendix A).

[14]Such as a bad harvest or oil price rises.

ows. Projection estimates of *GTC*, *TD*, nd *TC* are made on the basis of the usual ssumptions presented above.

II. Empirical Results

A. *Aggregate Trade Flows*

The estimated parameter values for the ›ooled regression for 1954–70 is:

$$(4)\quad \log X_{ij} = 6.72 + \underset{(.03)}{.788} \log Y_j^N + \underset{(.03)}{.954} \log Y_i^N - \underset{(.04)}{.177} \log N_j - \underset{(.04)}{.283} \log N_i - \underset{(.03)}{1.229} \log D_{ij} + \underset{(.10)}{2.788} \log P_{ij}$$

$R^2 = .58$; standard errors are shown in parentheses. All coefficients are statistically significant at the 0.01 level.

The coefficients of the trade-flow equation for the 17-year period confirm the expected theoretical pattern. In fact, our regression results and in particular an R^2 of .58 show that we have a respectable fit.[15] Based on Quandt's maximum likelihood technique and likelihood ratio test, I conclude that two breaks, and not one, occurred. A maximum maximorum is reached in 1964 and another local maximum in 1958.[16]

The earlier break in 1958 points out that the break with the Stalinist development program after 1954 and the attempts made towards joint planning did in fact create a new structure. Yet this new structure cannot be viewed as representing a state of integration. In fact, it represents a period where a policy of autarky was abandoned in favor of a policy where international trade would play a greater role. The structural break in 1964 can, however, be attributed to the beginning of integration.[17]

Given the above results, I decided to recalculate the trade-flow equation for a stable period between the two breaks. The years chosen for this period were 1960–64. This recalculated equation, when the dummy variable for *CMEA* membership is removed, should represent a stable preintegration structure. Based on this equation we estimate inter- and intra-*CMEA* trade for the years 1965–70.

The pooled equation for 1960–64 is:

$$(5)\quad \log X_{ij} = 8.574 + \underset{(.08)}{.580} \log Y_j^N + \underset{(.08)}{.910} \log Y_i^N + \underset{(.08)}{.111} \log N_j - \underset{(.08)}{.178} \log N_i - \underset{(.05)}{1.509} \log D_{ij}$$

$R^2 = .52$; standard errors are shown in parentheses. All coefficients except for N_j are significant at the 0.01 level. The population elasticity for N_j is not significantly different from zero.[18]

In Table 1 the *GTC*, *TD*, and *TC* effects of *CMEA* integration as well as the total trade figures for *CMEA* are presented. Note that since economic integration is presumed to be a cumulative process, one should find estimates of trade creation increasing from year to year with no reversals. In fact, our results in the case of *CMEA* as a whole confirm these expectations. Yet despite the increase in total trade and the existence of trade creation, an examination of the *TD* figures points out that with the exception of 1970, the *CMEA* member

[15] The R^2 in our regression is as high as the one found by Linnemann. The best R^2 achieved by Linnemann in a worldwide sample for 1958–60 was 0.64. A further comparison of these results with those of Linnemann, Glejser, and Pulliainen confirms both the direction and magnitude of the income and population elasticities.

[16] See the author (1976a, pp. 106–08).

[17] In the post-Stalinist period a great deal of planning towards integrated development took place. However, most of the recommendations remained just that. It was only after the 1962 "Basic Principles" that one notices some partial changes in intra-*CMEA* economic cooperation. For further discussion of the *CMEA* see the author (1976a, ch. 3).

[18] Linnemann, p. 15, attributes the negative influence of N on economies of scale in production. Leamer and Stern on the other hand, p. 153, suggest that the negative influence of N may be explained by opportunity cost theory. Typical estimates find the population elasticities around −.2 and not always significantly different from zero—consistent with our findings.

TABLE 1—NET EFFECT OF *CMEA* INTEGRATION ON *CMEA*
(Millions of *U.S* Dollars)

Year	*GTC*	*TD*	*TC*	Total Trade[a]
1965	9202.87	−769.41	9972.28	16496.98
1966	9235.08	−856.85	10091.93	17292.62
1967	10299.95	−904.39	11204.34	18986.17
1968	11263.27	−813.18	12076.41	20607.39
1969	12071.51	−429.95	12501.46	22481.53
1970	13221.59	122.34	13099.25	24853.69

Source: Trade flows between *CMEA* partners were provided by the Indiana University *IDRC* and *ITIMS* data bank. For a list of other sources, see the author (1976a, pp. 99–102).
[a]Total exports of the *CMEA* member countries to the countries in the sample.

countries continued to trade outside their customs union.

An evaluation of the individual country results[19] confirms both our a priori expectations and the results shown in Table 1. Moreover, these individual country estimates demonstrate that the only *CMEA* member countries which are effectively diverting trade from nonpartner to partner countries are Czechoslovakia and East Germany.

B. *Disaggregate Trade Flows*

The empirical analysis of the disaggregate trade flows was in general hampered by the lack of consistent reporting of disaggregate foreign trade flows by the *CMEA* countries. In fact, the Soviet Union, Czechoslovakia, and to a lesser extent, Poland, are the only members of *CMEA* which report a more or less consistent set of foreign trade statistics by commodity composition and partner.

While 37 disaggregated commodity groups were tested, the results for 34 of these commodity groups showed that only the preference variable was significantly different from zero. A number of reasons may have accounted for these results. First, in a large number of these 34 commodities, the influence of Soviet *CMEA* member bilateral trade flows was very strong. This was confirmed when an eighth independent variable, the total Soviet trade in each commodity group, was found to be significantly different from zero. Secondly, extra-*CMEA* transactions in some groups, i.e., beverages and tobacco, mineral fuels, lubricants and related materials, are so small compared to the intra-*CMEA* flows that they may be considered inconsequential.

The empirical results for the remaining three commodities, basic chemicals, iron and steel, and machinery other than electric were in fact consistent with the expected theoretical pattern and the aggregate results presented above. In the case of basic chemicals, these empirical results illustrate a structural break occurring in 1964. Using the recalculated trade-flow equation (6) for the last preintegration year, we obtain estimates of *GTC*, *TD*, and *TC* effects of *CMEA*.

The recalculated equation for 1964 is:[20]

$$(6)\quad \log X_{ij} = -.84 - \underset{(.11)}{.083 \log Y_j^N} - \underset{(.11)}{.01 \log Y_i^N} + \underset{(.12)}{.239 \log N_i} + \underset{(.12)}{.15 \log N_i} - \underset{(.07)}{.323 \log D_{ij}}$$

Standard errors are shown in parentheses

[19]For individual country estimates both for aggregate and disaggregate results see the author (1976b, pp. 11–18).

[20]R^2 coefficients for equations (6), (7), and (8) are not reported because an interpretable R^2 when using generalized least squares (*GLS*) estimation does not exist. Zellner's estimation used in the analysis of disaggregate trade flows is simply the application of *GLS* estimation to a group of seemingly unrelated equations. Furthermore, as Fisher points out "the orthogonality properties of least squares which makes R^2 easy to interpret in terms of fraction of variance are not preserved" (p. 34) in the above case. In order to appraise our results the standard errors of the coefficients are provided.

TABLE 2—NET EFFECT OF *CMEA* INTEGRATION ON *CMEA* IN BASIC CHEMICALS
(Millions of *U.S.* Dollars)

Year	*GTC*	*TD*	*TC*	Total Trade[a]
1965	167.39	-118.50	285.89	306.02
1966	163.65	-147.71	311.36	325.42
1967	180.23	-155.79	336.02	354.73
1968	203.63	-159.68	363.31	421.79
1969	206.41	-185.03	391.44	409.92
1970	227.75	-223.16	450.91	469.29

ee Table 1.

In Table 2 the *GTC*, *TD*, and *TC* effects *CMEA* integration on the trade flows of isic chemicals are presented. Note that the timates of trade creation are increasing om year to year despite the existence of ter-*CMEA* trade in basic chemicals.

An evaluation of the individual country timates for this commodity group conrms the results presented in Table 2, as ell as the aggregate results. While the *MEA* members did experience trade creaon for this commodity during 1965–70, iter-*CMEA* trade flows were still in exisnce.

The results for iron and steel, like those r basic chemicals, were significant. The mpirical results again support the hypothes of a structural break occurring in 1964. n this case the recalculated 1964 equation i:

$$\text{(}) \quad \log X_{ij} = -1.221 - \underset{(.11)}{.183} \log Y_j^N + \underset{(.11)}{.013} \log Y_i^N + \underset{(.12)}{.418} \log N_j + \underset{(.12)}{.311} \log N_i - \underset{(.08)}{.373} \log D_{ij}$$

Standard errors are shown in parentheses.

The projected trade flows based on this recalculated equation with the dummy variable for *CMEA* membership removed are used to determine the estimates of the *GTC*, *TD* and *TC* effects of *CMEA*. The results presented in Table 3 again confirm our expectations of positive trade creation. However, despite the existence of trade creation, the growth of inter-*CMEA* trade flows should be noted. In fact, an examination of the individual country estimates illustrates that in both Hungary and East Germany, inter-*CMEA* trade in iron and steel is larger than intra-*CMEA* trade in the same commodity.

In the case of machinery other than electric, the results of Quandt's maximum likelihood technique point to a structural break in 1962. Because this break comes so soon after the signing of the "Basic Principles" it suggests that this commodity group may be of greater importance to the *CMEA* members' industrialization drive.

Using the recalculated trade flow equation for 1962, we estimate inter- and intra-*CMEA* trade for years 1963–70. The re-

TABLE 3—NET EFFECT OF *CMEA* INTEGRATION ON *CMEA* IN IRON AND STEEL
(Millions of *U S* Dollars)

Year	*GTC*	*TD*	*TC*	Total Trade[a]
1965	881.33	-287.96	1169.29	1190.35
1966	821.41	-339.49	1160.22	1182.00
1967	883.02	-315.67	1198.69	1219.98
1968	900.88	-339.23	1240.11	1260.80
1969	1134.35	-410.35	1545.22	1565.77
1970	1260.31	-495.36	1755.67	1776.33

[a]See Table 1.

TABLE 4—NET EFFECTS OF *CMEA* INTEGRATION ON *CMEA* IN MACHINERY OTHER THAN ELECTRIC
(Millions of *U.S.* Dollars)

Year	*GTC*	*TD*	*TC*	Total Trade[a]
1963	1680.88	-95.96	1776.84	1816.99
1964	1817.37	-141.79	1959.16	1998.81
1965	1898 19	-128 08	2026.27	2069.12
1966	1889 08	-182.12	2071.20	2115.58
1967	2039.18	-185.40	2224.58	2270.98
1968	2231.23	-209.98	2441 21	2489.15
1969	2283.12	-209.28	2492.40	2541.84
1970	2626 70	-252 72	2879 42	2931.17

[a]See Table 1

calculated equation for 1962 is:

$$(8) \quad \log X_{ij} = -.114 + \underset{(.11)}{.228} \log Y_j^N + \underset{(.11)}{.2} \log Y_i^N + \underset{(.13)}{.131} \log N_j + \underset{(.13)}{.002} \log N_i - \underset{(.09)}{.447} \log D_{ij}$$

Standard errors are shown in parentheses.

The projected *GTC*, *TD*, and *TC* effects are presented in Table 4. The results here again confirm our a priori assumption of trade creation increasing from year to year with no reversals. In fact, compared with the total trade figures the dollar value of trade creation is quite large. With the exception of East Germany and Czechoslovakia, the other *CMEA* member countries have in fact diverted trade in this commodity from nonpartner sources to partner countries.

III. Concluding Remarks

The empirical results are found to be consistent with my expectations and those presented by the theory. The results with respect to aggregate trade flows showed that the *CMEA* countries have in fact experienced a cumulative growth in *GTC* and *TC* over the integration period 1965–70. The projected estimates of the size of the *GTC* effect ranged from $9.2 billion in 1965 to $13.2 billion in 1970. The estimates of *TC* ranged from $9.9 billion in 1965 to $13.1 billion in 1970.

In the case of dissaggregate trade flows, our results were not generally consistent with the aggregate results. The greatest discrepancy arises when comparing the size of *GTC* and *TC* between the aggregate and disaggregate results. A reasonable explanation for this may be found in the lack and suitability of disaggregated *CMEA* trade flows. In fact, in those commodity groups where the data were both available and consistent with the *SITC* nomenclature, the results were somewhat consistent with the aggregate results.

In basic chemicals the size of *GTC* grew from $167 million in 1965 to $228 million in 1970. The value of *TC* grew from $285 million in 1965 to $450 million in 1970. The size of *GTC* and *TC* effects in iron and steel also expanded from $881 million and $1169 million in 1965 to $1260 million and $1755 million in 1970, respectively. Finally, the projected estimates of the size of the *GTC* and *TC* effects in machinery other than electric ranged from $1.6 billion and $1.7 billion in 1963 to $2.6 billion and $2.9 billion in 1970, respectively.

My empirical findings yielded the additionally important conclusion that integration did, in fact, occur after the signing of the "Basic Principles." Moreover, these results demonstrate that 1964 can be considered the last preintegration year in *CMEA*.

REFERENCES

N. D. Aitken, "The Effect of the EEC and EFTA on European Trade: A Temporal Cross-Section Analysis," *Amer. Econ.*

Rev., Dec. 1973, *63*, 881–92.
la Balassa, "Trade Creation and Trade Diversion in the EEC," *Econ. J.*, Mar. 1967, *77*, 1–21.
——, *European Economic Integration*, Amsterdam 1975.
T. Berend, "The Problem of Eastern European Economic Integration in a Historical Perspective," in Imre Vajda and Mihaly Simai, eds., *Foreign Trade in a Planned Economy*, London 1971, 1–28.
I. K. Carney, "Development in Trading Patterns in the Common Market and EFTA," *J. Amer. Statist. Assn.*, Dec. 1970, *65*, 1455–59.
M. Fisher, "Simultaneous Equation Estimation: The State of the Art," Inst. Defense Analysis seminar paper, July 1970.
amas Foldi, *Studies in International Economics*, Budapest 1966.
—— and Tibor Kiss, *Socialist World Market Prices*, Budapest 1969.
I. Glejser, "An Explanation of Differences in Trade-Product Ratios Among Countries," *Cah. Econ. Bruxelles*, 1968, No. 37, 5, 47–58.
—— and A. Dramais, "A Gravity Model of Interdependent Equations to Estimate Flow Creation and Diversion," *J. Reg. Sci.*, Dec. 1969, *9*, 439–49.
. M. Goldfeld and R. E. Quandt, "The Estimation of Structural Shifts by Switching Regressions," *Ann. Econ. Soc. Measure.*, Feb. 1973, *4*, 475–85.
ranklyn D. Holzman, *Foreign Trade Under Central Planning*, Cambridge 1974.
H. B. Junz and R. R. Rhomberg, "Prices and Export Performance of Industrial Countries 1953–1963," *Int. Monet. Fund Staff Pap.*, July 1965, *12*, 224–69.
Michael Kaser, *Comecon: Integration Problems of the Planned Economies*, 2d ed., New York 1967.
——, *Economic Development for Eastern Europe*, New York 1968.
Heinz Kohler, *Economic Integration in the Soviet Bloc with an East German Case Study*, New York 1965.
Mordechai E. Kreinin, "Trade Creation and Diversion by the EEC and EFTA," *Econ. Int.*, May 1969, *22*, 273–80.
——, *Trade Relations of the EEC: An Emperical Investigation*, New York 1974.
Edward E. Leamer and Robert M. Stern, *Quantitative International Economics*, Boston 1970, 145–70.
Hans Linnemann, *An Econometric Study of International Trade Flows*, Amsterdam 1966.
Paul Marer, *Soviet and East European Foreign Trade, 1946–1969: A Statistical Compendium and Guide*, Bloomington 1972.
J. Pelzman, (1976a) "Trade Integration in the Council of Mutual Economic Assistance: Creation and Diversion 1954–1970," unpublished doctoral dissertation, Boston College 1976.
——, (1976b) "Economic Integration in C.M.E.A.," work. paper no. 10, Univ. South Carolina, Nov. 1976.
——, (1976c) "Trade Integration in the Council of Mutual Economic Assistance: Creation and Diversion—1954–70," *ACES Bull.*, Fall 1976, *18*, 39–59.
P. Poyhonen, "A Tentative Model for the Volume of Trade Between Countries," *Weltwirtsch. Arch.*, Band 90, 1963, I, 93–100.
——, "Towards a General Theory of International Trade," *Ekon. Samfundets Tidskr.*, Aug. 1963, *17*, 69–77.
K. Pulliainen, "A World Trade Study: An Econometric Model of the Pattern of the Commodity Flows in International Trade in 1948–1960," *Ekon. Samfundets Tidskr.*, Aug. 1963, *17*, 78–91.
R. E. Quandt, "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *J. Amer. Statist. Assn.*, Dec. 1958, *53*, 873–80.
——, "A Further Approach to the Estimation of Switching Regressions," work. paper memo. no. 122., Princeton Univ., Mar. 1971.
——, "Tests of a Hypothesis that a Linear Regression System Obeys Two Separate Regimes," *J. Amer. Statist. Assn.*, June 1960, *55*, 324–30.
——, "A New Approach to Estimating Switching Regressions," *J. Amer. Statist. Assn.*, June 1972, *67*, 306–10.

Jan Tinbergen, *Shaping the World Economy*, New York 1962.

E. Truman, "The EEC: Trade Creation and Trade Diversion," *Yale Econ. Essays*, Spring 1969, *9*, 201–57.

Imre Vajda and Mihaly Simai, *Foreign Trade in a Planned Economy*, London 1971.

Peter J. D. Wiles, *Communist International Economics*, New York 1969.

J. Williamson and A. Bottrill, "The Impact of Customs Unions on Trade in Manufactures," *Oxford Econ. Pap.*, Nov. 1971, *23*, 323–51.

A. Zellner, "Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results," *J. Amer. Statist. Assn.*, Dec. 1963, *58*, 977–92.

——, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregate Bias," *J. Amer. Statist. Assn.*, June 1962, *57*, 348–68.

University of Indiana, International Development Research Center (*IDRC*), International Trade Information Management Systems (*ITIMS*), data bank, various years.

Two-Sector Aggregative Models and the Investment Demand Function

By GEOFFREY WOGLOM*

Dale Henderson and Thomas Sargent (hereafter H-S) and Y. C. Park have analyzed the effectiveness of monetary and fiscal policy in the two-sector analogue of James Tobin's dynamic aggregative model. Contrary to the assumptions of Tobin's model, fiscal policy can affect real income in the H-S model. However, the sign of the effect of fiscal policy on real income depends on a number of conditions relating to the parameters of the money demand function and the capital intensities in the two sectors. A somewhat more disturbing result is that the sign of the effect of fiscal policy changes as one changes, *ceteris paribus*, the assumption of which sector is the more capital intensive. The H-S results seem to imply that the analysis of the effectiveness of fiscal policy in the traditional *IS-LM* analysis is very sensitive to the assumption of a one-sector production technology.

It is important to realize, however, that the H-S model differs from *IS-LM* analysis in two ways. Besides assuming a two-sector production technology, the H-S model also assumes a perfect capital market, where the asset value of capital is always equal to reproduction cost. Sargent and Neil Wallace have analyzed a one-sector model with and without the perfect capital market assumption (without and with a disequilibrium investment demand function). They find that many of the strange results of Tobin's model are eliminated if there are costs of adjusting the capital stock. For example, if the costs of adjustment are large enough, expansionary monetary policy lowers the interest rate as in *IS-LM* analysis, contrary to Tobin's results. Also, the stability conditions of the model imply that with costs of adjusting the capital stock, fiscal policy affects income and interest rates in much the same way as in *IS-LM* analysis.

This paper alters the H-S model to allow for an investment demand function based on costs of adjustment.[1] In analyzing the comparative static results of this model one can determine whether the strange results of the H-S model are due to the assumption of a two-sector production technology or the assumption of a perfect market in existing capital goods.

The profit-maximizing subsystem in the H-S model can be solved to yield the price level and the marginal product of capital as functions of the relative price of investment (see the Appendix). A general equilibrium occurs when the consumption good, money,[2] and investment good markets are in equilibrium.

$$(1)\quad g_2 + C(Y - T) - Y_2(K, N(Y, P_1), P_1) = 0$$

*Boston College, and Banking Section, Board of Governors of the Federal Reserve System. I would like to thank Dale Henderson for his patient help; remaining errors are my own. The Federal Reserve System is in no way responsible for the conclusions of this paper

[1]The assumptions underlying firm behavior necessary for an investment demand function in the H-S model are similar to the assumptions underlying investment demand functions in *IS-LM* models. Firms are assumed to produce both consumption and investment goods. It is also assumed that resources can be shifted costlessly between the production of both goods. The firm cannot, however, alter its total capital stock without incurring costs of adjustment. The *IS-LM* models add the further assumption that the production technologies are identical for both goods, and usually, also assume that the expected marginal physical product of capital is a constant

[2]There are only two asset market equilibrium conditions in the H-S model. money and nonmoney. H-S assume that all nonmoney assets (bonds and equities) are perfect substitutes and bear the same rate of return. The balance sheet constraint allows them to analyze asset market equilibrium in terms of the equilibrium condition in only one market. I follow their procedure and take money market equilibrium to imply an equilibrium in all asset markets.

$$(2) \qquad m(Y,r) - M/P(P_1) = 0$$

$$(3) \quad g_1 + I(R(P_1) - r) - Y_1(K, N(Y, P_1), P_1) = 0$$

where g_2 is government demand for consumption goods, and

- $C(\ )$ is a simple consumption function, $1 \geq C' \geq 0$
- Y is national income in real terms, $Y = P_1 Y_1 + Y_2$
- P_1 is real price of investment
- Y_1 is profit-maximizing supply of investment goods
- Y_2 is profit-maximizing supply of consumption goods
- T is lump sum taxes
- k_2; k_1 are the capital-labor ratios in the consumption and investment good sectors, respectively .
- K is total capital stock
- $N(\ )$ relates total employment to Y and P_1; $\partial N/\partial Y > 0$; $\partial N/\partial P_1 < 0$ (see Appendix)
- $Y_2(\ )$ relates Y_2 to K, N and P_1; $dY_2/dP_1 < 0$; $\partial Y_2/\partial Y \lesseqgtr 0$ as $k_1 \lesseqgtr k_2$
- $m(\ )$ is the real demand for money; $\partial m/\partial Y > 0$; $\partial m/\partial r < 0$
- r is the interest rate
- M is the nominal money supply
- $P(\ )$ relates the nominal price of consumption goods to the real price of investment goods, given profit maximization; $P' \gtreqless 0$ as $k_1 \gtreqless k_2$ (see Appendix)
- $R(\ )$ relates the marginal physical product of capital in the investment good sector to P_1; $R' \gtreqless 0$ as $k_1 \gtreqless k_2$ (see Appendix)
- $Y_1(\ )$ relates Y_1 to K, N and P_1; $dY_1/dP_1 < 0$; $\partial Y_1/\partial Y \gtreqless 0$ as $k_1 \gtreqless k_2$
- I is net investment demand; $I(0) = 0$; $0 < I' < \infty$ ($I' \rightarrow \infty$ is the H-S model)
- g_1 is government demand for investment goods

In analyzing the comparative static properties of the model it is convenient to solve equation (1), so that Y is a function of P_1, g_2 and T.

$$(4) \qquad Y = Y(P_1, g_2, T)$$

where

$$\frac{\partial Y}{\partial P_1} \gtreqless 0, \quad \frac{\partial Y}{\partial g_2} \gtreqless 0, \quad \frac{\partial Y}{\partial T} \lesseqgtr 0 \text{ as } k_1 \gtreqless k_2$$

The comparative static properties of the model (2)–(4) are summarized in Table 1.[3] The following stability condition[4] was used in deriving the comparative static results in Table 1:

$$(5) \quad -I'\left[\frac{MP'}{P^2} + \frac{\partial m}{\partial Y}\frac{\partial Y}{\partial P_1}\right] - \frac{\partial m}{\partial r} \cdot \left[I'R' - Y_1 - \frac{\partial Y_1}{\partial Y}\frac{\partial Y}{\partial P_1}\frac{-dY_1}{dP_1}\right] \lesseqgtr 0, \quad \text{as } k_1 \lesseqgtr k_2$$

A striking result of Table 1 is that fiscal policy that works through the investment good market affects the endogenous variables in a way qualitatively identical to IS-LM analysis.[5] The effect of an increase in M on Y and nominal prices is unambiguously positive as in the H-S model. The ef-

[3]The results of Table 1 are derived from the following matrix equation

$$\begin{bmatrix} 1 & \frac{-\partial Y}{\partial P_1} & 0 \\ \frac{\partial m}{\partial Y} & MP'/P^2 & \frac{\partial m}{\partial r} \\ \frac{-\partial Y_1}{\partial Y} & I'R' - \frac{\partial Y_1}{\partial P_1} & -I' \end{bmatrix} \begin{bmatrix} dY \\ dP_1 \\ dr \end{bmatrix} = \begin{bmatrix} \frac{\partial Y}{\partial g_2} \; dg_2 \\ dM/P \\ -dg_1 \end{bmatrix}$$

[4]A graphical analysis of the comparative static properties of the model and of the stability conditions of the model is available from the author on request. The stability condition (5) is derived from an analysis of the following dynamic model:

$$D(P(P_1)P_1) = a[g_1 + I(R(P_1) - r) - Y_1(K, N(Y, P_1), P_1)]$$

$$Dr = b[m(Y, r) - M/P(P_1)]$$

where $a > 0$; $b > 0$; $Y = Y(P_1, g_2, T)$.

[5]The asymmetry between the effectiveness of fiscal policy working through the investment good market and fiscal policy working through the consumption good market results from the lack of an interest rate effect in the consumption good market. If one assumed that consumption decreases when the interest rate rises (perhaps through a wealth effect), the effect of an increase in g_1 on Y becomes ambiguous. However, the effect of an increase in M on Y also becomes ambiguous.

TABLE 1—SUMMARY OF RESULTS

	Increase in					
	g_1	M		g_2		
					$k_1 > k_2$	
Effect on		IS slope > 0	IS slope < 0	$k_2 > k_1$	$A > 0$	$A < 0$
Y	+	+	+	?	+	?
P_1P (therefore P)	+	+	+	+	+	−
r	+	+	−	+	?	?

ct of an increase in M on the interest rate epends on the following condition:

$$(6)\quad \frac{dr}{dM} < 0 \quad \text{if} \quad I'R' - Y_1 - \frac{\partial Y_1}{\partial Y}\frac{\partial Y}{\partial P_1} - \frac{dY_1}{dP_1} \lesseqgtr 0 \quad \text{as} \quad k_1 \gtrless k_2$$

The nominal prices of investment and onsumption move in the same (opposite) irection as P_1, when $k_1 > k_2$ ($k_2 > k_1$, see ppendix). Condition (6) requires that the otal impact of an increase in nominal rices lowers the excess demand for investment. The adjective total refers to the fact hat as P_1 varies, Y varies to keep the onsumption market in equilibrium. Thus (6) is equivalent to the statement that a rise n the price level must lower the excess lemand for goods. This condition is nalogous to a negatively sloped IS curve n a one-sector model.

The necessary condition for the effect of an increase in g_2 to increase Y differs from the stability condition because of the following terms in the stability condition:

$$(7)\quad A = \frac{\partial Y}{\partial P_1}\left[-\frac{\partial m}{\partial Y}I' + \frac{\partial m}{\partial r}\frac{\partial Y_1}{\partial Y}\right]$$

In the case where $k_2 > k_1$ the presence of A helps to fulfill the stability condition, therefore the stability conditions are of no help in determining the effect of increases in g_2 on Y. In the case where $k_1 > k_2$ the sign of A is ambiguous. When A is greater than zero, the stability conditions imply that an increase in g_2 raises Y.

In the cases where the sign of the effect of an increase in g_2 on Y is ambiguous, one must examine the parameters of the money and investment good market equilibrium conditions. It is possible to develop conditions similar to those used by H-S (their interest inelastic case and the interest elastic case conditions) to determine the sign. However, if one knew that the production transformation surface was relatively flat so that $\partial Y/\partial P_1$ was close to zero, then A would be close to zero and the stability conditions would imply that the effect of an increase in g_2 on Y is positive. Thus most of the strange results of the H-S model result from the perfect capital market assumption and this statement becomes stronger as the transformation surface becomes flatter.

APPENDIX

The H-S profit-maximizing conditions (their equations (19)–(21)) can be written as:

$$(A1)\quad f_1(k_1(R)) - f_1'(k_1(R))k_1 = W_0/PP_1$$

$$(A2)\quad f_2(k_2(P_1R)) - f_2'(k_2(P_1R))k_2 = W_0/P$$

where f_1, f_2 are the intensive production functions in the investment and consumption good sectors

$k_1(R)$ is defined implicitly by $f'(k_1) = R$ and similarly for $k_2(P_1R)$

W_0 is the fixed money wage

In (A1) and (A2) two of the three prices (R, P_1, P) must be treated as endogenous. We will choose to treat P and R as endogenous to the profit-maximizing subsystem.

This implies the following comparative static properties:

$$\begin{bmatrix} -k_1 & \frac{W_0}{P_1P^2} \\ -k_2P_1 & \frac{W_0}{P^2} \end{bmatrix} \begin{bmatrix} dR \\ dP \end{bmatrix} = \begin{bmatrix} \frac{-W_0}{P_1^2P} & dP_1 \\ k_2R & dP_1 \end{bmatrix}$$

Let the determinant of the matrix $= B = \frac{W_0}{P^2}(k_2 - k_1)$.

Therefore,

$$\frac{dP}{dP_1} = P' = \left(-k_1k_2R - \frac{W_0}{P_1P}k_2\right)\Big/B \gtrless 0 \quad \text{as} \quad k_1 \gtrless k_2$$

$$\frac{dR}{dP_1} = R' = \frac{-W_0}{P_1P^2}\left(\frac{W_0}{P_1P} + k_2R\right)\Big/B \gtrless 0 \quad \text{as} \quad k_1 \gtrless k_2$$

$$\frac{dP(P_1)P_1}{dP_1} = -\left(\frac{W_0}{P}k_1 + k_1k_2RP_1\right)\Big/B \gtrless 0 \quad \text{as} \quad k_1 \gtrless k_2$$

In addition it is helpful to recall the properties of factor-market equilibrium in a two-sector economy (see Duncan Foley and Miguel Sidrauski).

(A3) $$\frac{\partial Y_1}{\partial N} \gtrless 0 \quad \text{as} \quad k_2 \gtrless k_1$$

(A4) $$\frac{\partial Y_2}{\partial N} \gtrless 0 \quad \text{as} \quad k_2 \lesseqgtr k_1$$

(A5) $$\frac{\partial Y_1}{\partial P_1} > 0$$

(A6) $$\frac{\partial Y_2}{\partial P_1} < 0$$

With these results all that remains is to analyze the determinants of N. Changes in N affect the outputs Y_1 and Y_2, but do not affect the capital-labor ratios and therefore do not affect factor prices. Since $Y = P_1Y_1 + Y_2$, the above results imply a relationship between Y, N, and P_1. A graph of the production possibilities for two levels of employment in the case where $k_2 > k_1$ is given in Figure 1.

In Figure 1 the two transformation surfaces differ by the amount of total employment (Figures 1, 2, and 3 always depict the case where $k_2 > k_1$). I have also drawn two parallel lines with slope equal to minus P_1. The profit-maximizing subsystem (A1) and (A2) will correspond to a tangency of the relative price line with a transformation surface. Points 1 and 2 are both consistent with the same relative prices and, therefore, the value of P_1. The dashed line from 1 to 2 shows that there is a locus of pairs of Y_2 and Y_1 consistent with the same P_1. In Figure 1 real income is given by the vertical intercept of the relative price line. There is a range of values of real income consistent with subsystem (A1) and (A2) given any P_1. As one moves from point 1 to 2 total employment increases. Therefore, given a P_1 total employment is a positive function of real income.

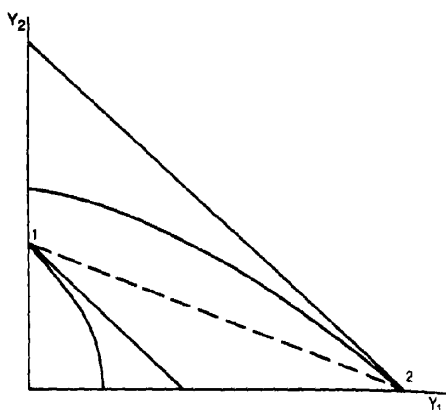
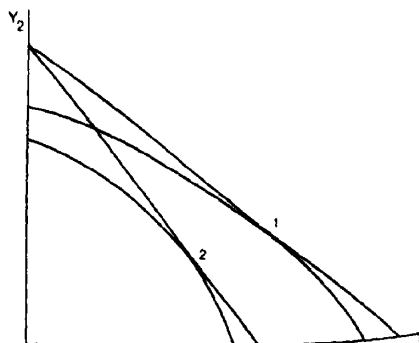



FIGURE 1

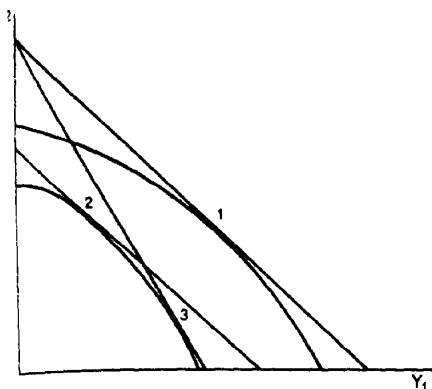



FIGURE 2

FIGURE 3

Figure 2 is helpful in analyzing how this relationship changes for changes in P_1. Figure 2 illustrates that if real income is to stay constant as P_1 rises (given the convexity of the transformation surfaces), total employment must fall.

$$(A7)\quad N = N(Y, P_1);\ \frac{\partial N}{\partial Y} > 0\ \ \frac{\partial N}{\partial P_1} < 0$$

At this point it is useful to analyze a source of ambiguity of the total effects of changes in P_1 on the supply of consumption goods in the case where $k_2 > k_1$ and Y is held constant. Given (A6) and (A7), the effect of a change in P_1 on the supply of consumption goods is given by the expression:

$$(A8)\quad \frac{dY_2}{dP_1} = \frac{\partial Y_2}{\partial N}\frac{\partial N}{\partial P_1} + \frac{\partial Y_2}{\partial P_1} \gtrless 0$$

If $k_2 > k_1$, the first term is positive. An increase in P_1 lowers total employment, which increases the output of the capital intensive sector. The second term, however, is negative. This ambiguity is somewhat similar to the ambiguity in consumer behavior of the effect of an own-price change on demand because of the income and substitution effects. Figure 3 presents a graphical analysis of the problem. The change in Y_2 from point 1 to point 2 corresponds to the first term in (A8) and the change in Y_2 from point 2 to point 3 corresponds to the second term.

In the H-S model this ambiguity is described in terms of the capital using effect (*KUE*) and the capital valuation effect (*KVE*). The *KUE* is analogous to the second term of (A8) and *KVE* to the first term. In their analysis H-S assume that k_1 is sufficiently close to k_2 so that *KUE* dominates *KVE*, or such that (A8) is negative. We will follow H-S and assume (A8) is negative.

REFERENCES

Duncan Foley and Miguel Sidrauski, *Monetary and Fiscal Policy in a Growing Economy*, London 1971.

D. Henderson and T. Sargent, "Monetary and Fiscal Policy in a Two-Sector Aggregative Model," *Amer. Econ. Rev.*, June 1973, *63*, 345–65.

Y. C. Park, "The Transmission Process and the Relative Effectiveness of Monetary and Fiscal Policy in a Two-Sector Neoclassical Model," *J. Money, Credit, Banking*, June 1973, *5*, 595–622.

T. Sargent and N. Wallace, "Market Transaction Costs, Asset Demand Functions, and the Relative Potency of Monetary and Fiscal Policy," *J. Money, Credit, Banking*, May 1971, *3*, 469–505.

J. Tobin, "A Dynamic Aggregative Model," *J. Polit. Econ.*, Apr. 1955, *53*, 103–15.

Some International Evidence on Output-Inflation Tradeoffs: Comment

By MARCELLE ARAK*

In the June 1973 issue of this *Review*, Robert E. Lucas presented a model for the determination of aggregate output and the price level which was based upon rational price expectations. His empirical results were generally quite favorable to the underlying theory. However, all of his estimates were based upon the assumption that nominal *GNP* was an exogenous variable. If this assumption is inappropriate, Lucas' method of testing his theory will yield biased results.

I develop new tests for his basic model which are valid even if nominal *GNP* is not exogenous. Applied to the United States, these tests do not support the underlying theory: prices are poorly explained by the equation based upon rationality, and errors in price expectations are not significantly associated with variations in output.

I. Derivation of the New Tests

Lucas assumes that nominal income does not depend upon the decisions made by suppliers of output; they merely decide how nominal *GNP* will be apportioned between output and prices. In the context of an aggregate demand function which relates quantity demand to the price level, this assumption requires that the price elasticity be unity—for only then would total expenditures be independent of the price level. My tests do not require that nominal *GNP* be exogenous in this sense and thus I need not assume any particular price elasticity.

Let ξ represent the price elasticity of demand for aggregate output. The demand curve is then written as:

$$(1) \qquad y_t = -\xi P_t + x_t$$

where y is the *log* of real *GNP*, P is the *log* of the price level, and x is an exogenous shift variable that has the same distributional properties as the x posited by Lucas, namely:

$$(2) \qquad x_t = x_{t-1} + \delta + U_t$$

where δ is a constant and U is a random variable which has expected value of zero and a variance of σ_U^2. According to equation (2), x typically rises by δ per period, but the stochastic element (U) with an expected value of zero may make x rise by more or less in any given time period. Although x has the same distributional properties as the x specified by Lucas, x in this model is not the equivalent of the *log* of nominal *GNP*. To see this, add P to both sides of the demand equation, (1):

$$(3) \qquad P + y = (1 - \xi)P + x$$

The left-hand side is the *log* of nominal *GNP* while the right-hand side is x only if $\xi = 1$. Therefore x and the *log* of nominal *GNP* will be identical only if $\xi = 1$.

The new demand equation (1) may be more conveniently cast in terms of the current unknown element U by taking the first difference of (1) and replacing $x_t - x_{t-1}$ by $\delta + U_t$:

$$y_t - y_{t-1} = -\xi \Delta P_t + \delta + U_t$$

or

$$(1') \qquad y_t = y_{t-1} + \xi P_{t-1} - \xi P_t + \delta + U_t$$

The supply equation is taken exactly specified by Lucas:

$$(4) \qquad y_t = y_{nt} + \theta\gamma(P_t - P_t^e) + \lambda(y_{t-1} - y_{n,t-1})$$

where P^e is the *log* of the expected p level and y_{nt} represents the *log* of the n mal level of real output at time t.

*Federal Reserve Bank of New York. The views expressed are those of the author and not necessarily those of the Federal Reserve Bank of New York. I would like to thank Michael Hamburger and Alan Spiro for helpful comments.

Following Lucas, I assume that the *log* of the price level will be a linear combination of past and present exogenous variables, past endogenous variables, and the current random element (U):

$$(5)\quad P_t = K_0 + a_0 U_t + \sum_0^{\infty} b_i y_{n,t-i} + \sum_0^{\infty} c_i y_{t-1-i} + \sum_0^{\infty} d_i P_{t-1-i} + \sum_0^{\infty} e_i Z_{t-i}$$

where the Z are any other predetermined variables that might be relevant. Rational price expectations then imply that

$$(6)\quad P_t^e = K_0 + \sum_0^{\infty} b_i y_{n,t-i} + \sum_0^{\infty} c_i y_{t-1-i} + \sum_0^{\infty} d_i P_{t-1-i} + \sum_0^{\infty} e_i Z_{t-i}$$

since $E(U_t) = 0$. The a_i, b_i, c_i, d_i, and e_i can be obtained by substituting for P and P^e in the demand and supply curves, setting demand equal to supply, and equating the coefficients of each variable. This yields

$$K_0 = \frac{\delta}{\xi} \qquad b_1 = \frac{\lambda}{\xi}$$
$$a_0 = \frac{1}{\theta\gamma + \xi} \qquad c_0 = \frac{(1-\lambda)}{\xi}$$
$$b_0 = -\frac{1}{\xi} \qquad d_0 = 1$$

and all other $a_i, b_i, c_i, d_i, e_i = 0$.

Substituting these parameter values into equation (5) and replacing y_{nt} by $y_{n,t-1} + \beta$, the price level can be expressed as

$$(7)\quad P_t = P_{t-1} + \frac{(1-\lambda)}{\xi}(y_{t-1} - y_{n,t-1}) + \frac{\delta - \beta}{\xi} + \frac{U_t}{\theta\gamma + \xi}$$

where β is the growth rate of normal output. To obtain the reduced form for output, note that $P_t - P_t^e = a_0 U_t = U_t/(\theta\gamma + \xi)$. The supply equation can therefore be written as

$$(8)\quad y_t = y_{nt} + \lambda(y_{t-1} - y_{n,t-1}) + \frac{\theta\gamma}{\theta\gamma + \xi} U_t$$

Equations (7) and (8) which are the reduced forms for the price level and output can be used to test the model and derive its parameters. The λ is obtained by estimating the coefficient of $(y_{t-1} - y_{n,t-1})$ in equation (8); ξ is obtained by estimating the coefficient of $(y_{t-1} - y_{n,t-1})$ in equation (7), $(1-\lambda)/\xi$, and using the λ from equation (8). The $\theta\gamma$ is obtained by regressing the computed residuals from (8), $\theta\gamma U_t/(\theta\gamma + \xi)$, upon the computed residuals from (7), $U_t/(\theta\gamma + \xi)$.

Lucas estimates

$$(9)\quad y_t - y_{nt} = -\pi\delta + \lambda(y_{t-1} - y_{n,t-1}) + \pi\Delta G + \epsilon_t$$

where G is the *log* of nominal *GNP*, π is $\theta\gamma/(\theta\gamma + 1)$, and ϵ_t is an error term assumed to be $N(0, \sigma_\epsilon^2)$. Whether the underlying model is correct or not, the estimate of $\theta\gamma$ implied by the coefficient of ΔG is likely to be biased if ξ is not exactly equal to unity. If the underlying model is correct, but $\xi \neq 1$, the error term contains $-\pi(1-\xi)\Delta P$.[1] A comparison of the reduced form for P (equation (7)) with the reduced form for G (equation (7) plus (8)) indicates that a positive correlation exists between G and P. As a result, the coefficient of ΔG would tend to be biased upward (downward) if ξ is greater (less) than unity; the estimate of $\theta\gamma$ would be biased in the same direction as π. If the underlying model is not correct or is only approximately correct, ϵ will contain other things. But if ϵ contains other things which have an influence on the supply of real output and a

[1]Note that equation (3) implies that the x is related to the *log* of nominal *GNP*: $x = G - (1-\xi)P$. Thus U which equals $x_t - x_{t-1} - \delta$ can be expressed as $U = \Delta G - (1-\xi)\Delta P - \delta$. Substituting this into equation (8), we obtain:

$$y_t - y_{nt} = -\frac{\theta\gamma}{\theta\gamma + \xi}\delta + \lambda(y_{t-1} - y_{n,t-1}) + \frac{\theta\gamma}{\theta\gamma + \xi}\Delta G - \frac{\theta\gamma}{\theta\gamma + \xi}(1-\xi)\Delta P$$

shock to the supply function has an effect on nominal GNP, then ϵ will indirectly have an influence on G. Thus the coefficient of G will be biased upward (downward) if a positive shock to real output tends to raise (lower) nominal GNP. This means that by the Lucas method one could obtain a positive and significant $\theta\gamma$ even if output did not in fact respond to errors in price expectations.

II. Results

The model was tested for the United States for the years 1952–67. (The National Income Accounts data that were used reflect the benchmark revisions of January 1976, although similar results are obtained from the original data.) G is the *log* of nominal GNP; y is the *log* of real GNP, P is the *log* of the GNP deflator; and the *log* of normal output, y_n, is derived from fitting a time trend.[2] If the Lucas output equation is estimated using the revised data, we obtain:

$$y_t - y_{nt} = \underset{(.004)}{.045} + \underset{(.058)}{.885}(y_{t-1} - y_{n,t-1}) + \underset{(.067)}{.837}\Delta G_t$$

$$R^2 = .962; S.E.E. = .0064; D.W. = 0.75$$

where standard errors are noted below each coefficient. The coefficient of ΔG represents $\theta\gamma/(\theta\gamma + 1)$ assuming the elasticity of demand is unity. Accordingly, $\theta\gamma$ is about 5—each percentage point error in the price expectation leads to a 5 percent expansion in supply, *ceteris paribus*—and $\theta\gamma$ is significantly greater than zero at the 95 percent confidence level.[3]

Turning to the new estimates, a very different picture emerges. The three estimated equations that provide the tests of the model are:

$$(10)\quad y_t - y_{nt} = \underset{(.19)}{.74}(y_{t-1} - y_{n,t-1}) + V_t$$

$$R^2 = .50;\quad S.E.E. = .021;\quad D.W. = 2.07$$

$$(11)\quad P_t - P_{t-1} = \underset{(.002)}{.020} + \underset{(.065)}{.089}\cdot(y_{t-1} - y_{n,t-1}) + W_t$$

$$R^2 = .12;\quad S.E.E. = .0074;\quad D.W. = 1.21$$

$$(12)\quad V_t = \underset{(.74)}{.86}\,W_t$$

$$R^2 = .08;\quad S.E.E. = .020;\quad D.W. = 1.77$$

where the V_t and W_t are estimated residuals from the output equation (with an intercept of zero) and the price equation, respectively. They imply the price elasticity of demand ξ is 2.9 and the supply response $\theta\gamma$ is .86. These values are quite different from the Lucas parameters. More important, however, these estimated equations yield a very different view of the adequacy of the underlying model. Equation (11), the price equation implied by the model, does not significantly explain movements in the price level. And equation (12) indicates that the supply response $\theta\gamma$ is not significantly different from zero. Although the reduced form for output is significant, it is merely a test of whether output is autocorrelated. In sum, these empirical results call into question the basic model postulated by Lucas. Certainly for the United States it appears that either the supply function, the demand function, or the assumption of rationality must be altered.

[2]The *log* of the normal level of output was estimated for 1951–67 as:

$$y_{nt} = \underset{(0.02)}{6.30} + \underset{(.002)}{.034}t \quad S.E.E. = .031$$

where t = 1 in 1951.

[3]$\theta\gamma = \pi/(1-\pi)$ if $\xi = 1$. Thus the variance of the estimated $\theta\gamma$ is

$$E\left(\frac{\pi}{1-\pi} - \frac{\hat{\pi}}{1-\hat{\pi}}\right)^2$$

The asymptotic standard error of $\theta\gamma$ is equal to (the standard error of $\pi)/(1-\pi)^2$ or about 2.5.

REFERENCE

R. E. Lucas, "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

Some International Evidence on Output-Inflation Tradeoffs: Reply

By ROBERT E. LUCAS, JR.*

Marcelle Arak's comment develops from the observation that *if* the solutions (11) and (12) of my 1973 model (p. 329) were *exact* (that is, if their error terms were identically zero), then the aggregate supply elasticity can be identified from the covariance matrix of the residuals. In this case, my unmotivated assumption that the demand elasticity is unity would be unnecessary for identification of the supply elasticity. In other words, if aggregate demand shifts were to trace price and output movements along a *perfectly* stable supply curve (or Phillips curve), then one could estimate this curve without any prior restrictions on the slope of the demand curve.

In my paper, I did not assume an exact fit in estimating the parameter π, though I did say that "the fits should be 'good'" (p. 330). The quotation marks on good were intended as an admission of an inability to make this restriction precise; because of this imprecision, the goodness of fit criteria was used only informally in evaluating the model, by reference to "high" R^2s.

In the Arak version, an exact fit *is* assumed, which leads to a very sharp test: her model implies that the error terms V_t and W_t in her (10) and (11) should be *perfectly correlated*, since both are proportional to the aggregate demand disturbance U_t. In fact, their squared correlation is .08 (Arak, equation (12)). Hence, the exact version of the model must be rejected. It follows that the supply response $\theta\gamma$ cannot be inferred from (10)-(12).

In short, the data reject Arak's identifying restriction on the supply function error. What about my restriction that $\xi = 1$? Her estimate is 2.9, calculated as the ratio $(1 - .74)/.089$; the numerator has a standard error of .19, the denominator of .065. These numbers cannot reject $\xi = 1$ very decisively. Nor is the goodness of fit of (10) and (11) evidence against my model: if (as I assumed) cyclical movements result mainly from demand shifts and if (as Arak assumes) demand shifts are unobservable shocks entering the error terms of (10) and (11), then the theory *predicts* poor fits on these equations! The evidence that *does* hurt is the $R^2 = .08$ in (12): I would have calculated this number had I thought the econometrics through as clearly as Arak has, and I would have expected it to be "near" unity.

So as not to lose sight of the objectives of my 1973 study, let me insist that the issues raised in the Arak paper have no more bearing on "the assumption of rationality" than they do on, say, the assumption of utility maximization. The question is only whether a particular, rather arbitrary, two-variable model provides a good enough approximation to observed behavior to support the international Phillips curve comparisons I wanted to make. My strategy was based on the hope that the sample would exhibit enough across-country variation to overcome what must necessarily be large measurement errors on the supply elasticities for each country individually. The fact that the main prediction of the natural rate hypothesis was confirmed suggests that this strategy was successful. Had it *not* been confirmed (as would have been the case if no South American countries had been included) I would, of course, have concluded that my two-variable model was inadequate, not that agents are "irrational."

*University of Chicago.

REFERENCES

M. Arak, "Some International Evidence on Output-Inflation Tradeoffs: Comment," *Amer. Econ. Rev.*, Sept. 1977, *67*, 728–30.

R. E. Lucas, "Some International Evidence On Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

The Price Equation: A Cross-Sectional Approach

By RONALD P. WILDER, C. GLYN WILLIAMS, and DAVINDER SINGH*

The imposition of price controls in the *U.S.* economy during the period 1971–73, together with the persistent inflationary bias in the decade of the 1970's, has brought forth a renewed research interest in the process of price determination.[1] The long-standing debate regarding whether firms tend to use full-cost or target return pricing rules on the one hand, or marginalist principles on the other has been intermingled with questions regarding the efficacy of price controls. In addition, the administered-price inflation hypothesis has been subjected to renewed empirical testing. Each of these areas of inquiry is of current policy relevance to the general problem of inflation.

Although this paper will touch on each of the above areas, its principal objectives are narrow:

1. to specify a cross-sectional price equation which is consistent with both target return and marginalist principles;
2. to estimate the parameters of the price equation using data on year to year rates of change during the period 1958–72, for a large sample of four-digit *SIC* manufacturing industries;
3. to test hypotheses: a) regarding the relative effects of cost changes, demand changes, degree of concentration and price expectations on industry prices; and b) regarding the effects of the Phase I and II price controls of 1971–72.

Many studies have examined the issues discussed in this paper. A discussion of selected studies which have treated some or all of the issues listed above will illustrate their current status. Most researchers agree that cost changes show up more strongly in price equations than do demand changes, although that result could be because the latter seem to be more difficult to measure. Researchers also agree that there are differences in the price adjustment process among concentrated and nonconcentrated industries, but the exact nature of these differences is controversial. The issue of rule of thumb versus marginalist pricing practices has also not been settled in the empirical literature since most price equations are consistent with either practice.

In their study of the price equation, Eckstein and Fromm test the relative significance of target return and competitive elements in the determination of price changes. The equations which they estimate cover all manufacturing, durable manufacturing, and nondurable manufacturing based on two-digit *SIC* sectors. They utilize a time-series approach with data mainly for the period 1954:1 to 1965:4. Cost and demand elements are included in the equations. The authors make the general conclusion that: "While the different forms of the equations yield varying results on the relative importance of the competitive mechanism vis-à-vis oligopolistic pricing, there is pretty strong evidence that equations combining both mechanisms are superior to equations using either approach in isolation" (p. 1171).

In a cross-section study of 395 four-digit *SIC* manufacturing industries, Frank Ripley and Lydia Segal estimate a price change equation for the period 1959–69. Changes in unit labor cost, changes in the materials cost, changes in real output, and changes in productivity are independent variables, with industries classified by broad concentration category. They find that over the eleven year period covered by their study concen-

*Department of economics, University of South Carolina.

[1] Rather than cite each item in the extensive literature, the following articles and their bibliographies provide broad coverage: William Nordhaus, Otto Eckstein and Gary Fromm, and Eckstein and David Wyss on price equations; Robert Lanzillotti, M. Hamilton and Blaine Roberts on Phase II controls; J. Fred Weston on alternative models of pricing behavior; and Steven Lustgarten and Ralph Beals on tests of the administered pricing inflation hypothesis. A more general survey on inflation has recently been provided by David Laidler and Michael Parkin.

trated industries passed on into prices a smaller proportion of unit labor cost than did nonconcentrated industries (p.268). This leads them to suggest that an incomes policy which controlled wage and price movements discriminatorily by industry should be designed to control prices in concentrated industries and wages in nonconcentrated industries.

Two recent studies summarize the work done in estimating the impact of the level of concentration in product markets on the timing and magnitude of industry price changes. Beals' review of the literature for the Council on Wage and Price Stability and Lustgarten's study for the American Enterprise Institute both conclude that prices behave differently in more than in less concentrated industries. Beal's work emphasizes the greater stability of prices associated with degree of concentration (p.44). Lustgarten (1975a) emphasizes the tendency of prices in concentrated industries to rise less than in nonconcentrated industries because of the more rapid productivity growth in the former (p.36). Lustgarten's study is of special interest in its relevance to the findings recorded in this paper since he uses a data base similar to that used here.

The present study, while using a data base similar to that used by Ripley and Segal, and Lustgarten (1975a), differs from those studies by virtue of its emphasis on the broader questions of the price equation relationships and by its examination of changes over time in the cross-sectional price equation.

I. The Price Equation: Specification

Following Eckstein and Fromm we first develop a price equation incorporating marginalist principles. Suppose the competitive firm maximizes short-run gross returns to capital:

$$\pi = pq - wL(q) - P_m mq \tag{1}$$

where q is the rate of output, p is the output price, w the wage rate, $L(q)$ total manhours, P_m material prices, and m material inputs per unit of output. Profit maximization requires that:

$$p = w(dL/dq) + P_m m \tag{2}$$

For discrete price changes, when the change in marginal cost equals the change in unit variable cost (and deleting second-order interactions)

$$\Delta p = w\Delta(L/q) + (L/q)\Delta w + P_m \Delta m + m\Delta P_m \tag{3}$$

Hence, the change in price is the sum of changes in unit labor cost (ULC) and unit materials cost (UMC), or the change in variable cost per unit (VC):

$$\Delta p = \Delta ULC + \Delta UMC = \Delta VC \tag{4}$$

For a firm operating as a monopolist or in monopolistic competition, and using optimal markup pricing, since $p = MR(1/1 - 1/e) = MC(1/1 - 1/e)$ when the profits are maximized (letting MR and MC denote marginal revenue and marginal cost, and e denote the price elasticity of demand), equation (4) becomes

$$\Delta p = \Delta VC\left(1/1 - \frac{1}{e}\right) \tag{5}$$

assuming constant elasticity of demand.

The other major variant of the price equation is target rate of return pricing. Again following Eckstein and Fromm:

$$p = \frac{\bar{\pi}K}{\bar{q}} + \overline{VC} \tag{6}$$

where $\bar{\pi}$ is target rate of return on the firm's capital stock K, $\bar{q}$ is standard rate of output, and $\overline{VC}$ is unit variable cost at the standard rate of output. For discrete price change:

$$\Delta p = \frac{K}{(\bar{q})}\overline{\Delta\pi} + \overline{\pi\Delta}\frac{K}{(\bar{q})} + \Delta\overline{VC} \tag{7}$$

If target rate of return $\bar{\pi}$, capital stock K, and standard volume $\bar{q}$ are relatively stable over the interval measured, changes in unit variable cost should predominate in the price equation.

Changes in unit variable cost are prominent in each of the price equations discussed here. Demand changes are also important, with elasticity changes affecting the profit-maximizing price and hence the optimal markup over variable unit costs. Hence, our

specification of the price equation for empirical testing takes the following form:

$$(8) \quad \Delta p = \alpha_0 + \alpha_1 \Delta VC + \alpha_2 \Delta IS + a_3 CONC + U$$

where ΔIS is a demand proxy (described below), $CONC$ is a concentration measure, and U is the error term. Since our data source is a sequence of cross sections of year-to-year rates of change in price, costs, and other variables, we are also able to include an industry concentration measure ($CONC$) in the price equation to test for differences in pricing behavior as between more and less concentrated industries.

This specification of the price equation is consistent with either marginalist or cost markup pricing principles and therefore cannot discriminate between the two,

II. Empirical Tests: 1958–72

The data set is based on the *Annual Survey of Manufactures* and on the industry price indexes developed by the Bureau of Economic Analysis, Department of Commerce, from the wholesale price indexes of the Bureau of Labor Statistics.[2] The sample includes 357 four-digit industries for most of the years during the period. The data are annual averages, and hence the cross sections of year-to-year changes represent first differences of annual average values.

A. *Variables*

Price (Δp), the dependent variable in each equation, is the annual percentage rate of change in the industry wholesale price index, computed between successive years.

Unit Variable Costs (ΔVC) are the annual percentage rates of change of the sum of unit labor costs and unit material costs.[3]

TABLE 1—SAMPLE MEANS AND STANDARD DEVIATIONS, FOUR-DIGIT MANUFACTURING INDUSTRIES (Annual Percentage Rates)

Year	ΔP	ΔVC	ΔIS
1958-59	1.47	1.20	42.62
	(3.85)	(5.04)	(853.17)
1959-60	.68	.08	3.73
	(3.37)	(5.41)	(48.01)
1960-61	−0.19	1.44	7.31
	(3.39)	(7.68)	(83.98)
1961-62	0.08	−0.1	32
	(2.65)	(4 19)	(50 60)
1962-63	0 29	−1.04	13.71
	(3 66)	(6.03)	(227 69)
1963-64	0 58	.33	38.70
	(3.37)	(4.11)	(504 12)
1964 65	1 10	.96	− 72
	(3 33)	(4.96)	(13 72)
1965 66	2.44	2.57	2.39
	(3.47)	(4.54)	(12.77)
1966-67	1.37	35	33 26
	(3 75)	(5.01)	(5 85)
1967-68	3.02	2.36	.48
	(3.88)	(5.22)	(10.24)
1968-69	3 23	3.35	4 52
	(4 00)	(5 58)	(48 08)
1969-70	3.72	3.98	968.25
	(4.41)	(5.95)	(1921 21)
1970-71	3 12	1.85	593 18
	(5.16)	(5.93)	(6196.57)
1971 72	3.65	3.38	−6.04
	(5 31)	(8 08)	(24 47)

Note: Standard deviations are in parentheses

Inventory Sales Ratio (ΔIS) is a demand proxy: the annual percentage rate of change of the ratio of an industry's average inventory to its value of shipments.[4] A large positive change in this ratio is indicative of slack demand, while a near zero or negative

[2]The price data were obtained on tape from the Division of Research and Statistics of the Federal Reserve Board. This data set was used earlier by Ripley and Segal.

[3]Unit labor costs and material costs were computed by defining an index of real output $X = VS/P$, where VS is the value of industry shipments and P is the industry price index. Unit labor cost is defined: $ULC = PAY/X$ where PAY is total payroll, while unit material cost is defined $UMC = MC/X$ where MC is total material cost.

[4]We experimented with the following alternative specifications of the demand variable: a) An inventory shipments ratio based on the change relative to the three previous years and on the change relative to a five-year moving average; and b) A capital expenditures measure, based on the change relative to the five-year moving average. Each of these alternative specifications yielded results similar to those reported. Use of capacity utilization or new orders variables as demand proxies was not possible due to the non-availability of data at the four-digit level.

TABLE 2—REGRESSION RESULTS FOR PRICE EQUATION, FOUR-DIGIT MANUFACTURING INDUSTRIES

Year	Coefficients for: Intercept	ΔVC	ΔIS	$CONC$	R^2
1958-59	.80[a]	.59[a]	.0001	-.0002	.55
1959-60	.54	.36[a]	-.002	-.003	.34
1960-61	.14	.21[a]	-.002	-.006	.23
1961-62	1.04[a]	.45[a]	.003	-.02	.47
1962-63	.84[a]	.50[a]	.0006	-.001	.50
1963-64	.27	.61[a]	-.0001	.002	.55
1964-65	.99[a]	.51[a]	-.04[a]	-.007	.60
1965-66	1.87[a]	.63[a]	-.03[a]	-.02[a]	.61
1966-67	1.18[a]	.56[a]	-.0003	-.002	.50
1967-68	2.27[a]	.58[a]	-.06[a]	-.01[a]	.58
1968-69	2.66[a]	.52[a]	-.01	-.02[a]	.55
1969-70	1.88[a]	.55[a]	-.00001	-.005	.52
1970-71	1.26[a]	.70[a]	-.00002	.01	.65
1971-72	1.99[a]	.58[a]	-.01	-.01	.63

[a]Significant at the 5 percent level.

change in the ratio is associated with a high level of demand. Hence, the postulated sign for this variable in the price equation is negative.

Concentration Ratio ($CONC$) is the 8-firm value of shipments concentration ratio from the 1967 *Census of Manufactures*. The administered-price inflation hypothesis would suggest a positive sign for this variable.

B. *Results*

Sample means and standard deviations are reported in Table 1. The cross-sectional distribution of annual rates of price change has experienced an increase in absolute dispersion as well as in mean value during the period 1958-72. The distribution of rates of change in the inventory variable (ΔIS) has moved with the business cycle as one would expect, with the largest mean value occurring in the 1969-70 recessionary period and the smallest mean value occurring during the 1971-72 recovery period.

The basic price equation tested, (8), was fitted to the annual percentage rates of change for each successive pair of years during 1958-72. It relates rates of change in price to rates of change in unit variable costs, the inventory shipments ratio, and to the level of the 8-firm concentration ratio; with the results shown in Table 2. Linear forms and ordinary least squares estimating procedures are used throughout.

The estimated regression coefficients for the three independent variables, ΔVC, ΔIS, and $CONC$, are reasonably stable throughout the study period. The coefficient on ΔVC is generally in the range of .5 to .6 and is of course highly significant. There is no evidence from these results of an increase in the markup over unit variable costs when the price equation is interpreted in this fashion. The major departures of the coefficient on unit variable costs from the .5 to .6 range are in 1959-60 and 1960-61, a period characterized by a business cycle peak in April 1960 and a trough in February 1961, and in 1970-71, a period containing a trough in November 1970 and the Phase I price freeze of August 1971. It should be noted that a coefficient on ΔVC of below unity does not necessarily imply that total dollar profits or rates of return are decreasing over time.[5]

[5]With fixed costs constant, the change in total profits $\Delta\pi$ may be expressed as follows: $\Delta\pi = \Delta P - \Delta VC$ where $\Delta P = \alpha_0 + \alpha_1 \Delta VC$ as in the linear regression. Then $\Delta\pi = \alpha_0 + \alpha_1 \Delta VC - \Delta VC$ or $\Delta\pi = \alpha_0 + \Delta VC(\alpha_1 - 1)$. In order that $\Delta\pi > 0$, it is necessary that $\alpha_1 > 1 - \alpha_0/\Delta VC$, a condition which is satisfied for most of the time periods in Table 2.

The demand variable (ΔIS) has the expected negative sign in the majority of cases, but is statistically significant only during the boom years of 1964–65, 1965–66, and 1967–68. These relatively weak results for the demand variables are consistent with the finding in the Nordhaus survey of pricing studies that "...demand variables do not show up consistently and significantly" (p.41).

Our estimated coefficients for the concentration variable are not consistent with the administered-price inflation hypothesis. The coefficients are negative in all years but one, and are statistically significant (negative) in three years. These results are consistent with those of the recent studies by Lustgarten (1975a,b).

The most interesting result shown in Table 2 is the upward trend in the intercept term, beginning in the mid-1960's. Over the period 1958–59 to 1964–65, the intercept term averaged less than 1 percent, while in the period 1965–66 to 1971–72, the average estimated annual rate of price change in the absence of change in the independent variables was closer to 2 percent. These results reflect the inflationary bias of the late 1960's and early 1970's and suggest that the inflationary tendency is not a result of changes in cost markup factors.[6]

Suspecting that the upward drift in the intercept term might reflect changes in inflationary expectations or in lags in adjustment to cost changes, we respecified the price equation to include lagged values of the rates of change in unit variable costs (ΔVC_{t-1}) and price (ΔP_{t-1}). The inclusion of the lagged price variable is based on the hypothesis that inflationary expectations are based on recent experience in price movements.

The results of estimating the coefficients for this expanded price equation are shown in Table 3. The lagged cost variable (ΔVC_{t-1}) has the expected positive sign. The lagged price variable (ΔP_{t-1}), however, is negative in the majority of cases. We had expected positive signs for the lagged price variable on the grounds that it is a measure of price expectations and that industries experiencing more rapid rates of price change in the past would be more inclined to raise price in the current period. The negative sign, while not consistently significant, would tend to suggest the existence of price rigidity in the sense that high rates of price change in the previous period are associated with lower rates of change in the current period. The inclusion of the lagged variables has a negligible effect on the intercept terms and other estimated coefficients relative to the values reported in Table 2.

A further development of the annual cross-sectional model is to consider the question of whether prices respond differently to decreases in variable costs as compared to increases in variable costs. There have been numerous allegations in the literature that prices tend to be relatively inflexible downwards even in the face of decreases in unit variable costs. To test this hypothesis, a cost increase/cost decrease dummy variable (VI) was added to equation (8). This variable takes on the value 1 for industries for which $\Delta VC \geq 0$ (increase in unit variable costs) and 0 otherwise. The estimated regression coefficients for this version of the price equation are shown in Table 4.

The coefficients for the cost change dummy variable are negative in twelve of the fourteen time periods, with seven of the negative coefficients statistically significant. These results provide fairly strong evidence of an asymmetry in the response of price to changes in unit variable costs. The nature of this asymmetry is that a larger relative decrease in the rate of

[6]The upward trend in the intercept term could be a result of a trend in unit fixed costs. To shed some light on this possibility, we computed total fixed costs and profits as follows: $TFC = VS - PAY - MC$ where TFC = total fixed costs and profits, PAY = total payroll, and MC = total materials costs. Over the period 1958–72 the aggregate proportion of fixed costs and profits (as defined here) to value of shipments increased from .21 to .25. However, when we regressed the annual intercept terms in Table 2 on the annual aggregate means of unit fixed costs and profits, the regression coefficient was statistically nonsignificant, with a t value of less than 1.0. Hence, the role of fixed costs in explaining the upward trend in the intercept is not straightforward.

TABLE 3—REGRESSION RESULTS FOR PRICE EQUATION WITH LAGGED COST AND PRICE VARIABLES

Year	Coefficients for:						
	Intercept	ΔVC_t	ΔIS_t	$CONC$	ΔVC_{t-1}	ΔP_{t-1}	R^2
1959–60	.64	.37[a]	−.002	−.004	.04	−.09	.35
1960–61	.09	.32[a]	−.002	−.006	−.27[a]	−.28[a]	.30
1961–62	1.03[a]	.45[a]	.0004	−.02[a]	.04[a]	.01	.48
1962–63	.77[a]	.50[a]	.0006	.0004	.08	.01	.50
1963–64	.32	.60[a]	−.0001	.001	.02	−.06	.54
1964–65	.78[a]	.56[a]	−.03[a]	.003	.30[a]	−.26[a]	.64
1965–66	1.59[a]	.58[a]	−.02[a]	−.01[a]	.05	.13[a]	.63
1966–67	1.47[a]	.58[a]	−.0004	−.01	.18[a]	−.22[a]	.52
1967–68	2.24[a]	.58[a]	−.06[a]	−.01[a]	−.001	.03	.58
1968–69	1.98[a]	.55[a]	−.005	−.01[a]	.26[a]	−.09	.61
1969–70	2.24[a]	.56[a]	−.0001	−.01	.06	−.12[a]	.51
1970–71	1.92[a]	.72[a]	.002	.01	.16[a]	−.33[a]	.66
1971–72	1.96[a]	.59[a]	−.01	−.01	10[a]	−.10[a]	.63

[a]Significant at the 5 percent level

change of unit variable costs is required to obtain a given *reduction* in the rate of price change as compared to the increase in unit variable costs needed to obtain a similar *increase* in the rate of price change. The results therefore are consistent with some degree of downward price rigidity.

The final development in estimating the price equation is to pool the annual cross sections, using dummy variables to differentiate among the annual periods. Interaction variables between concentration and unit variable costs and concentration and the demand proxy are introduced. The estimating equation for the pooled cross sections is as follows:

$$(9)\quad \Delta P = \alpha_0 + \alpha_1 \Delta VC + \alpha_2 \Delta IS + \alpha_3 CONC + \alpha_4 CONC \cdot \Delta VC + \alpha_5 V1 + \alpha_6 CONC \cdot \Delta IS + \sum_{2}^{n} \alpha_{5+j} T_j + U$$

where the T_j are the n-1 dummy variables

TABLE 4—REGRESSION RESULTS FOR PRICE EQUATION, WITH COST INCREASE/COST DECREASE DUMMY VARIABLE

Year	Coefficients for:					
	Intercept	ΔVC	ΔIS	$CONC$	Cost Change Dummy ($V1$)	R^2
1958–59	1.73[a]	.68[a]	.00003	−.004	−1.47[a]	.55
1959–60	.01	.30[a]	−.001	−.004	1.01[a]	.36
1960–61	−1.24[a]	13[a]	−.002	−.002	2.58[a]	34
1961–62	1 03[a]	46[a]	0006	−.02	−.10	.48
1962–63	1 57[a]	.58[a]	.0007	−.004	−.33[a]	.51
1963–64	.74	.73[a]	−.00002	.004	−1 10[a]	.62
1964–65	1 51[a]	.59[a]	−.03[a]	−.006	−1 17[a]	61
1965–66	2.52[a]	.67[a]	−.03[a]	−.02[a]	−.98[a]	.60
1966–67	1 67[a]	.69[a]	−.0004	−.003	−.66	.50
1967–68	2.39[a]	.59[a]	−.06[a]	−01[a]	−18	.58
1968–69	2.96[a]	.53[a]	−.006[a]	−.02[a]	−.42	.50
1969–70	1.76[a]	.53[a]	−.00001	−.006	−.26	.50
1970–71	2.33[a]	.77[a]	.00003	.006	−1 46[a]	.63
1971–72	3.20[a]	.63[a]	−.01	−.10[a]	−1.56[a]	.64

[a]Significant at the 5 percent level

TABLE 5—ESTIMATED PRICE EQUATION, POOLED CROSS SECTIONS

	Total Pooled Sample Cost-Change Dummy Variable: Excluded (1)	Included (2)	Pooled Sample Subsets with: $\Delta VC \geq 0$ (3)	$\Delta VC < 0$ (4)
Intercept	.94[a]	1.06[a]	.28	1.53[a]
ΔVC	.69[a]	.70[a]	.77[a]	.72[a]
ΔIS	.00001	.00001	.00001	-.0002
$CONC$	-.002	-.002	.008[a]	-.007[a]
$CONC \cdot \Delta VC$	-.32[a]	-.33[a]	-.48[a]	-.31
$CONC \cdot \Delta IS$	-.00004	-.00005	-.0001	.0008
$V1$	-	-.21[a]	-	-
T2 (1959-60)	-.60[a]	-.59[a]	-.63[a]	-.56
T3 (1960-61)	-1.05[a]	-.11[a]	-1.05[a]	-1.11[a]
T4 (1961-62)	-.72[a]	-.72[a]	-.76[a]	-.74[a]
T5 (1962-63)	-.03[a]	-.05	-.33	.08
T6 (1963-64)	-.47[a]	-.47[a]	-.25	-.79[a]
T7 (1964-65)	-.25	-.26	-.41	-.07
T8 (1965-66)	.31	.33	.43	.05
T9 (1966-67)	.38	.39	.47	.22
T10 (1967-68)	.92[a]	.93[a]	1.07[a]	.64
T11 (1968-69)	.71[a]	.72[a]	.78[a]	.63
T12 (1969-70)	.98[a]	.99[a]	1.18[a]	.47
T13 (1970-71)	1.27[a]	1.27[a]	1.63[a]	.64
T14 (1971-72)	.83[a]	.83[a]	.88[a]	.87[a]
R^2	.56	.56	.47	.35
Sample Size	5284	5284	3342	1942

[a]Significant at the 5 percent level

(n = 14) corresponding to the yearly intervals, with 1958 as the omitted category.

Equation (9) was estimated for the total pooled sample (with and without the cost dummy variable $V1$) and for the subsets defined to include those observations with $\Delta VC < 0$ and $\Delta VC \geq 0$, respectively. The pooled cross-section results are shown in Table 5. Sample size for the pooled sample is in excess of 5000 observations.

Again the asymmetry in price response to cost decreases as compared to cost increases is evident, as shown by the negative coefficient for $V1$ in column (2) in Table 5 and the difference between the intercept terms in columns (3) and (4). The positive and statistically significant coefficients for the time dummy variables T10 through T14 are also consistent with the upward drift in the intercept terms in Tables 2 and 3. The interaction between industry concentration and unit variable cost ($CONC \cdot \Delta VC$) is negative and statistically significant, which suggests that concentrated industries pass a smaller proportion of unit variable cost changes than do less concentrated industries, a finding consistent with that of Ripley and Segal. This interaction appears to be the principal way in which concentrated industries differ from less concentrated industries with respect to pricing behavior. In general, the remaining results for the pooled cross sections are similar to those for the individual annual cross sections.

III. Empirical Tests: The Controls Period of 1971-72

The consensus of most studies of th[e] Phase I and Phase II controls of 1971-72 [is] that controls had a short-run impact in r[e]ducing the rate of price change at the a[g]gregate final consumption level, but ha[d] little or no effect in the manufacturing se[c]tor.[7] All of the previous studies of the co[ntrols]

[7]See, for example, Lanzillotti and Roberts, [...] Lanzillotti, Hamilton, and Roberts.

TABLE 6—REGRESSION RESULTS, PHASE II PRICE CONTROLS ERA

Year	Intercept	ΔVC_t	ΔIS_t	$CONC$	ΔVC_{t-1}	Δp_{t-1}	T_1	T_2	R^2
1970-71a	1.23[a]	.70[a]	.0001	.01			−.08	.41	.65
1970-71b	1.70[a]	.73[a]	−.0001	.01	.17[a]	−.31[a]	−.20	.33	.68
1971-72a	2.06[a]	.58[a]	−.01	−.01			−.19	−.36	.63
1971-72b	2.02[a]	.59[a]	−.01	−.01	−.10[a]	−.10[a]	−.17	−.35	.63

[a]Significant at the 5 percent level.

trols period of which we are aware, however, utilize time-series data. We believe that the relatively disaggregated cross-sectional data set utilized here offers additional insight into the effects of controls and avoids the problem of separating the business cycle effects on labor productivity and unit labor costs associated with the recovery from the 1970 recession from the effects of controls on prices during the 1971-72 period.

Our test of the efficacy of price controls employs the alternative specifications of the price equation as reflected in Tables 2 and 3, with the addition of two price controls dummy variables based on whether an industry was characterized by firms fitting into the respective Phase II categories: Tier I, Tier II, or Tier III.[8] Our hypothesis is that the Tier I and Tier II dummy variables should have a negative sign given the more intensive prenotification and/or reporting requirements to which firms in these categories were subjected relative to the smaller Tier III firms.

The estimated coefficients for the basic price equations, modified by including the Tier I and Tier II dummy variables (Tier III is the omitted category), are reported in Table 6.[9] The Tier variables were not expected to be statistically significant in 1970-71, since Phase II controls did not go into effect until November 14, 1971. For the 1971-72 period, the Tier variables are negative as hypothesized, but not statistically significant (t-values were less than 1.0 in both specifications). Moreover, the magnitude of the Tier II dummy variable is larger than that for Tier I, which is the reverse of our expectations. Based on the coefficients for the equation labeled 1971-72a in Table 6, industries for which the leading firms are in the Tier I category experienced annual rates of price change only 0.2 percentage points lower than the relatively unregulated Tier III industries, a difference not statistically significant.

IV. Summary and Conclusions

This study has utilized a data set of fourteen annual cross sections of rates of change in price, cost, and a demand proxy to test various hypotheses regarding the price equation in the manufacturing sector. The primary findings are the following:

1) The coefficient on unit variable costs has been reasonably stable over the fourteen-year period, which is suggestive of stability in markup pricing factors.

2) The findings on the effect of industry concentration are not consistent with the administered-price inflation hypothesis. In general, concentrated industries pass along a smaller proportion of unit variable cost changes into price changes than do less concentrated industries. Therefore, for given rates of cost increase, the effect of concentration on the rate of price increase is weakly negative.

3) The inflationary tendency of the late 1960's and early 1970's in $U.S.$ manufactur-

[8]The Tier dummy variables are constructed by computing the average size (in value of shipments) of the eight largest firms by multiplying total industry value of shipments by the 8-firm concentration ratio and dividing the result by 8. Hence the Tier I dummy is 1 for an industry which has an average firm size of greater than $100 million, and 0 otherwise; the Tier II dummy is 1 for average firm sizes between $50 and $100 million, and 0 otherwise. For details of Phase II enforcement, see Lanzillotti and Roberts.

[9]The equations in Table 6 were also estimated with the inclusion of the cost change dummy variable $V1$, with virtually identical results for the Tier variables.

ing is not explained satisfactorily by increased rates of change in unit variable costs, increases in the markup on unit variable costs, or by changing demand conditions.

4) The findings suggest an asymmetry in the response of price increases and decreases in unit variable costs, which is consistent with downward price inflexibility.

5) The findings on the short-run effects of Phase II price controls in manufacturing in 1971-72 are consistent with the negative findings of previous studies.

REFERENCES

R. Beals, "Concentrated Industries, Administered Prices and Inflation: A Survey of Recent Empirical Research," paper prepared for the Council on Wage and Price Stability, June 17, 1975.

O. Eckstein and G. Fromm, "The Price Equation," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1159-83.

——— **and D. Wyss,** "Industry Price Equations," in Otto Eckstein, ed., *The Econometrics of Price Determination, Conference*, Washington 1972.

D. Laidler and M. Parkin, "Inflation: A Survey," *Econ. J.*, Dec. 1975, *85*, 741-809.

Robert Lanzillotti and Blaine Roberts, "The Legacy of Phase II Price Controls," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 82-87.

———, ———, **and M. Hamilton,** *Phase II in Review: The Lessons and Legacy of Price Control*, Washington 1974.

Steven Lustgarten, (1975a) *Industrial Concentration and Inflation*, Washington 1975.

———, (1975b) "Administered Inflation: A Reappraisal," *Econ. Inquiry*, June 1975, *13*, 191-206.

W. Nordhaus, "Recent Developments in Price Dynamics," in Otto Eckstein, ed., *The Econometrics of Price Determination. Conference*, Washington 1972.

F. Ripley and L. Segal, "Price Determination in 395 Manufacturing Industries," *Rev Econ. Statist.*, Aug. 1973, *55*, 263-71.

J. F. Weston, "Pricing Behavior of Large Firms," *Western Econ. J.*, Mar. 1972, *10*, 1-18.

U.S. Bureau of the Census, *Annual Survey of Manufactures*, Washington, various years.

———, *Census of Manufactures, 1967 Concentration Ratios in Manufacturing*, Washington 1970.

Unemployment, Inflation, and Monetarism: A Further Analysis

By ROBERT VAN ORDER*

With the acceleration of the controversy between Monetarists and Keynesians, it has become apparent that there is great need for a rigorous analysis of the dynamics of competing macroeconomic models. An important contribution to this area is a recent paper by Jerome Stein. My paper is intended to present a dynamic analysis of some of the issues involved in monetarist models of inflation and unemployment in a simplified version of the Stein model. The justification for adding what follows is that Stein's specific analysis of the dynamics of the system is for some purposes unnecessarily complicated.

What follows takes the Stein model as a point of departure. The dynamics of that model ultimately come from three adjustment equations: one giving wage inflation as a function of unemployment and expected inflation; one giving price inflation as a function of wage inflation and the excess demand for goods; and one giving the adjustment of expected inflation. The problem with getting a clear picture of how the entire system adjusts is that it is difficult to get a firm grip on three differential equations at once. The central purpose of this paper is to eliminate one of the equations. This is done by assuming that prices clear goods markets instantaneously. That is, prices are given by equating aggregate demand and supply, the tradeoff between inflation and unemployment coming from unemployment affecting wages, which affect equilibrium prices.

There are two reasons for making the assumption. First, while it is not unrealistic to argue that goods markets do not clear immediately, it does not appear to serve a major purpose to assume they do not clear. It is necessary to have labor markets out of equilibrium to have unemployment, and it is necessary to have expectations adjust to obtain the "dynamic Phillips curve" consistent with recent experience. However, once these assumptions are made, further disequilibrium assumptions (realistic though they may be) only cloud the analysis.

There is a second, more substantive theoretical issue. The model, of course, is meant to depict actual disequilibrium transactions. Unemployment is the difference between the usual labor demand and supply curves, coming from high real wages. Yet during every instant of the adjustment we could observe an excess supply of goods, i.e., overproduction. We should expect, given that we are talking about actual transactions, that this continual excess supply should eventually affect actual employment directly even if real wages remain constant. There are two ways out: one is to relate unemployment (and also price and wage adjustment) to excess supplies in both markets.[1] The other is to assume away the issue by having goods markets always clear. Simplicity being a virtue, the latter way out is taken here.

Two further simplifying assumptions are made in what follows: first, no underlying growth is assumed (i.e., the system is essentially stationary); the second, the interesting distinction that Stein makes between government purchases of goods and of labor services is ignored. The government purchases goods, produces nothing itself, and hires no labor.

*Economist, U.S. Department of Housing and Urban Development. The paper was written at the University of Southern California. I am grateful to an anonymous referee for helpful comments

[1]An approach to this issue is given in the author. On the direct effect of goods markets on employment see Robert Barro and Herschel Grossman.

I. The Model

In the discussion, the following symbols are used:

- g, g^* actual and expected rates of inflation
- y real output
- f indicator of fiscal policy
- M, m nominal and real money supply
- P, W levels of prices and wages
- V velocity of circulation
- n growth rate of M
- a_{ij} elements of matrix
- λ speed of adjustment of expected inflation
- θ ratio of bonds to money

Functions:

- P Phillips curve
- F production function
- E aggregate demand
- S aggregate supply
- R reduced form of E and S

Given the modifications in the introduction, the analysis follows Stein fairly closely. The labor market adjusts via

$$(1) \qquad g = P(y) + g^* \qquad P' > 0$$

where g and g^* are actual and expected rates of wage change and y is real output. Expectations adjust via

$$(2) \qquad Dg^* = \lambda(g - g^*) \qquad \lambda > 0$$

where the operator D represents differentiation with respect to time. This implies

$$(3) \qquad Dg^* = \lambda P(y)$$

The goods market is given by the usual aggregate demand and supply curves. The latter is given by

$$(4) \qquad \frac{P}{W} = S(y)$$

where P and W are prices and wages. Aggregate demand is given by

$$(5) \qquad \frac{P}{W} = \frac{M}{W} \cdot E(y, f, g^*, \theta)$$

or

$$(5') \qquad \frac{Py}{M} = yE(y, f, g^*, \theta) = V$$

where M is the money supply, f a measure of fiscal policy, and θ the ratio of private financial wealth to money (see Stein, p. 874). Equation (5′) defines velocity and indicates that it depends on fiscal policy even if long-run output is independent of f (or M). Hence, f can affect nominal income even in the long run.

Either (5) or (5′) can be viewed as derived from an elaboration of the standard textbook income expenditure model. The effects of y, f, and g^* on E are assumed to be positive, but the sign of θ is ambiguous since a rise in bond holdings tends to raise aggregate demand via a wealth effect and to lower it by increasing the demand for money, leading to "crowding out."

The key simplification made here is that prices instantaneously equate the right-hand sides of (4) and (5). The solution for y then is given by

$$(6) \qquad y = R(f, m, \theta, g^*)$$

with

$$m = \frac{M}{W}$$

Then

$$(7) \qquad Dy = R_f Df + R_m Dm + R_\theta D\theta + R_{g^*} Dg^*$$

with

$$(7') \qquad Dm = m(n - g)$$

where n is the rate of growth of the money supply.

At this point it is necessary to discuss what is exogenous and what is not. As Carl Christ, and Alan Blinder and Robert Solow (1973) have pointed out, the government faces a budget constraint that requires that its purchases plus interest payments be financed by taxes plus bond sales plus money creation. That is, the government has four instruments: expenditures, taxes, bond sales, and money creation, but only the time paths of three of them can be chosen exogenously. If for instance expenditures, taxes, and money are set, bond holdings are endogenous and θ cannot be taken as exogenous. This will be discussed below. Here we keep with the standard as-

sumption (as in Stein's model) that M and θ are exogenous. This means that changes in f are not arbitrary, but of a "balanced budget" type. Assume f, n, and θ are constant. Then from (1) and (2), (7) becomes

$$Dy = -R_m m[P(y) + g^*] + R_{g^*}\lambda P(y) \tag{8}$$

Equations (3) and (8), two differential equations in y and g^*, govern the motion of the system with fixed monetary and fiscal policies. The main point of the analysis is that this two-dimensional system is quite manageable and yields some insights into macroadjustments. The first issue, however, is the characterization of some comparative statics.

In the long run the time derivatives are zero, leaving a rather simple solution. From (3) real income is given by the Phillips curve and is independent of aggregate demand,[2] and from (7′) the rate of inflation (which is perfectly anticipated) equals the rate of monetary growth. From (5′) it is clear that fiscal policy can affect the level of prices although in the long run it does not affect the rate of growth of prices.

In the very short run real income is variable and wage levels are predetermined. Income is given by (6), which again is an elaboration of standard textbook stuff. Effects of changes in f, m, g^*, and θ have already been discussed. The more important issue for present purposes is the "medium run" in which dynamics must be analyzed.

II. Dynamics

Assuming y and g^* are measured in deviations from their long-run levels, a linear approximation to (3) and (8) is given by

$$Dy = -[R_m mP' - R_{g^*}\lambda P']y - R_m mg^* \tag{9}$$

$$Dg^* = \lambda P'y \tag{10}$$

or

$$\begin{bmatrix} Dy \\ Dg^* \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & 0 \end{bmatrix} \begin{bmatrix} y \\ g^* \end{bmatrix} \tag{11}$$

We know that a_{21} is positive and a_{12} negative. Therefore, the condition that the adjustment be locally stable is that

$$a_{11} = -[R_m mP' - R_{g^*}\lambda P'] < 0 \tag{12}$$

which is not necessarily true. In particular the speed of adjustment of expectations λ can always be chosen large enough to make the system unstable. This is a fairly common result (see Miguel Sidrauski).[3]

Assume expectations do not adjust too rapidly, so that a_{11} is negative. Then the system is stable, and its motion is given by the roots of the matrix in (11). These are given by

$$2s = -[R_m mP' - R_{g^*}\lambda P'] \pm [(R_m mP' - R_{g^*}\lambda P')^2 - \lambda P' R_m m]^{1/2} \tag{13}$$

The adjustment is stable, but it may be cyclical. In particular, even if λ is small enough for stability, it may be large enough to make the adjustment cyclical.

The phase diagram of the adjustment in the nonspiral case is given in Figure 1. Equilibrium is given by the intersection of the lines denoted by $Dg^* = 0$ (from (3)) and $Dy = 0$ (from (8)). The motion in the four areas set off by these lines is given by the arrows at right angles to each other in the figure. The longer arrows give typical paths of y and g^*.

In the nonspiral case the adjustment path must approach AA' as it approaches equilibrium. AA' has a slope equal to the ratio of the components of the characteristic vector associated with the larger of the two roots. This reflects the long-run dominance of the larger root in the adjustment. It can easily be shown that AA' is downward slop-

[2] This need not be the case, of course, if the capital stock is endogenous.

[3] It should also be noted that the magnitude of the slope of the Phillips curve P' has nothing whatever to do with stability. The solution will, of course, be stable regardless of the size of λ if the effect of inflationary expectations on aggregate demand R_{g^*} is negative.

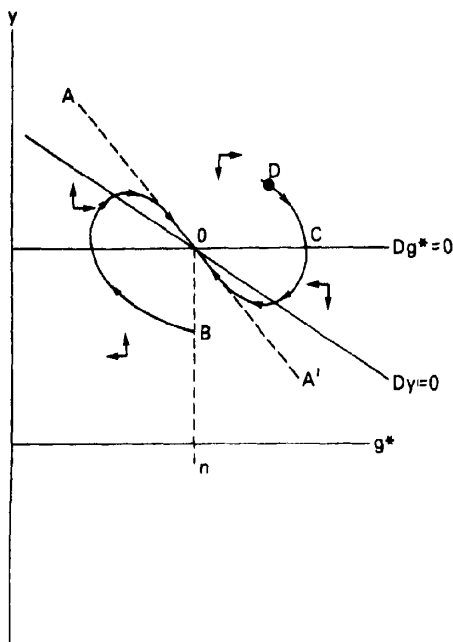



FIGURE 1 TIME PATHS OF OUTPUT AND EXPECTED INFLATION

ing but is steeper than the $Dy = 0$ line. Hence, adjustment paths must look like those depicted in the figure. In particular, as equilibrium is approached, expected wage inflation and income must be moving in opposite directions. In the cyclical case the paths are clockwise spirals. The motion of g is more complicated than that of g^*, but it can be deduced from (1) and (2), which imply that Dg depends on y and Dy. When y and g^* are both moving monotonically, g and g^* follow similar paths.

The model allows analysis of comparative dynamics. For simplicity the analysis is begun in long-run equilibrium. Consider first a once and for all technological change that leaves the natural unemployment rate constant, but raises production at that level. This raises both the $Dy = 0$ and $Dg^* = 0$ curves in the figure. However, as we know that the long-run inflation rate cannot be affected by this, we know both curves shift upward by the same amount, and the adjustment must follow a path like BO in the figure. That is, at first there is unemployment, then output rises, overshoots the long-run level and then declines. The inflation rate falls at first, and then goes back to its original level.

Consider a decline in the rate of monetary growth. This shifts the $Dg^* = 0$ line downward, leaving the long-run level of output unchanged. The assumption that we begin in equilibrium is maintained. Then the adjustment path looks like CO in the figure with output falling and then rising as the rate of inflation declines.

The simplified model brings out the issue of "stagflation," i.e., high inflation and high unemployment at the same time. Suppose that f is increased, raising aggregate demand. This raises both employment and inflation. Eventually we might expect f to be cut back perhaps to its initial level, putting us in a position like that at point D in the figure. We shall follow a path like DCO and throughout the latter stages of the adjustment (i.e., beyond the segment DC), both inflation and unemployment are greater than they had been initially.[4] Indeed, as was pointed out above, as equilibrium is approached unemployment and inflation must *both* be above equilibrium or below equilibrium, and they must be moving in the same direction.[5]

Finally, there is the issue of the differences between Monetarists and Keynesians. It is convenient to separate these issues into long run, short run and medium run, the first two relating to comparative statics in either the long run or short run and the last relating to comparative dynamics. The model in this paper has nothing new to say about the first two, which will be summarized briefly, but it does allow some insights into the last question.

Monetarists are inclined to emphasize

[4]Note that when f is decreased, putting the initial conditions in the northeast quadrant, expected inflation continues to worsen for a while even though the tight fiscal policy has lowered output. This brings out the lag with which inflation adjusts.

[5]This does not, of course, contradict the downward sloping short-run Phillips curve, since at this point it is shifts in the curve due to changes in expectation that are dominating the system.

the long run. In the long run, wages are perfectly flexible and inflation is perfectly anticipated. The equilibrium level of output is independent of both monetary and fiscal policy, and the rate of inflation equals the rate of growth of the money supply. In the short run, emphasized by Keynesians, employment is variable and can be affected in the usual ways by monetary and fiscal policy. It is not exactly the case as often argued by Monetarists (see Stein, p. 883) that output is independent of monetary growth and dependent only on the acceleration of monetary growth. In the short run, it is the difference between the rate of growth of the money supply and the inflation rate that matters. This is related to the rate of acceleration of the money supply, but is not the same. In the long run, of course, neither growth nor acceleration of money affects y.

Monetarists are inclined to argue that a policy (for example, change in f) that does not affect the money supply will have little effect on income. There are two basic reasons for this. The narrow one often associated with the Federal Reserve Bank of St. Louis argues that velocity is constant. The equation for velocity (5′) does not imply this, but of course whether or not the effect of f on V is "large" is an empirical question.

The second reason, brought out by Milton Friedman, concerns wealth effects. The argument is that a perpetual rise in government spending financed by selling bonds to the public will require a perpetual rise in the stock of government bonds held by the public, which will raise liquidity preference, raising interest rates and lowering y toward its original level. However, the rise in wealth also raises aggregate demand making the effect of θ on y ambiguous.[6] Strictly speaking the model here does not help answer the question. This is because it has been assumed that the bond supply is exogenous, and Friedman's argument makes it endogenous. This forces us to add an extra differential equation, giving the rate of change of bonds as a function of things in the government budget constraint. Blinder and Solow have pursued this in a short-run fixed price model and obtained a most interesting result: that if the effect of a rise in the bond supply is greater on liquidity preference than on the demand for goods, as Friedman's effect requires, then the long-run stock flow equilibrium is *unstable*. It is beyond the scope of this paper to follow this up in the more complicated dynamics of the model here, but this is clearly an issue worth pursuing.[7]

Medium-term dynamic issues are concerned with whether or not the system is stable and/or cyclical and, particularly in the stable noncyclical case, with the shapes of the paths. Equation (12) gives stability conditions. As mentioned above a large value of λ may lead to instability or to cyclical adjustment. Hence, strongly adaptive expectations complicate the adjustment. On the other hand a more "rational" type of expectation formation such as setting g^* equal to n with a small λ contributes to stability. It can also be seen from (12) that R_m enters into the dynamics. If, as Keynesians might argue, R_m is small, the system will tend toward instability and vice versa for Monetarists.

A central point of Monetarist analysis is that the $Dg^* = 0$ line is flat and the $Dy = 0$ line rather steep (its steepness is positively related to R_m, the effect of real balances on y). A Keynesian position might be that the $Dg^* = 0$ line is not flat, assuming that the coefficient of g^* in (1) is less than unity, but that is not of central importance. A significant Keynesian assumption would be that the $Dy = 0$ line is rather flat, being perfectly flat when $R_m = 0$ in which case prices rush violently toward zero or infinity depending on whether y is above or below the full-employment level and would have little effect on y.

Finally, the speed of adjustment of out-

[6] Arguing that the wealth effect of government bonds is cancelled out by expected future tax liabilities does not help, since the effect on liquidity preference should be cancelled out as well.

[7] For an exchange on the Blinder-Solow model, see Ettore Infante and Stein, and Blinder and Solow (1976).

put toward its equilibrium level can be seen from (9). For given initial levels of y and g^*, increases in R_m and P' and decreases in R_{g^*} will (in the stable case) speed up the adjustment of output. From this it is easy to see that the Keynesian emphasis on slow price adjustment (small P') and a small effect of monetary policy slows down the adjustment toward equilibrium and makes the short run more important, and vice versa for Monetarists. Thus, the simple model enables one to view more easily Monetarist vs. Keynesian arguments in dynamic terms; although, perhaps surprisingly, the crucial parameters turn out to be those emphasized in most static textbook models, R_{g^*}, R_m, and P'.

REFERENCES

R. Barro and H. Grossman, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, *39*, 17-26.

A. Blinder and R. Solow, "Does Fiscal Policy Really Matter?," *J. Publ. Econ.*, Nov. 1973, *2*, 310-37.

——— **and** ———, "Does Fiscal Policy Still Matter?," *J. Monet. Econ.*, Nov. 1976, *2*, 501-10.

C. Christ, "A Simple Macroeconomic Model with a Government Budget Constraint," *J. Polit. Econ.*, Jan./Feb. 1968, *76*, 53-67.

M. Friedman, "Comments on the Critics," *J. Polit. Econ.*, Sept./Oct. 1972, *80*, 906-50.

E. Infante and J. Stein, "Does Fiscal Policy Matter?," *J. Monet. Econ.*, Nov. 1976, *2*, 473-500.

M. Sidrauski, "Rational Choice and Patterns of Growth," *J. Polit. Econ.*, July/Aug. 1969, *77*, 575-85.

J. Stein, "Unemployment, Inflation, and Monetarism," *Amer. Econ. Rev.*, Dec. 1974, *64*, 867-87.

R. Van Order, "Excess Demand and Market Adjustment," *Econ. Inquiry*, Dec. 1976, *14*, 587-603.

Measurement of Tax Progressivity

By DANIEL B. SUITS*

Everyone knows that some taxes are progressive and others are regressive, and that there are degrees of each. Yet there is no generally accepted index of how progressive or regressive any given tax is. The purpose of this article is to present a widely useful index of tax progressivity, explore its properties, and to apply it to the analysis of the progressivity of a number of *U.S.* taxes. The index, inspired by and related to the Gini ratio, varies from +1 at the extreme of progressivity where the entire tax burden is borne by members of the highest income bracket, through 0 for a proportional tax, to −1 at the extreme of regressivity at which the entire tax burden is borne by members of the lowest income bracket. When the index is applied to 1970 data, the most highly progressive *U.S.* tax proves to be the federal corporate income tax with an index of +.32. The most regressive are sales and excise taxes with an index of −.15.

A useful property of the index is that the index of progressivity of a tax system consisting of two or more taxes is a weighted average of their individual indexes. On this basis, the entire 1970 *U.S.* tax system was very slightly progressive with an index of +.070.

Comparison of the 1970 values of the index with those of 1966 reveal that, although there was virtually no change in the progressivity of the *U.S.* tax structure as a whole, there were interesting changes in progressivity of individual taxes. In particular, the federal tax structure became somewhat more progressive, but the change was almost exactly offset by declining progressivity of state and local taxes.

*Professor of economics, Michigan State University. I am indebted to Joseph Pechman and Benjamin Okner for their helpful suggestions, and for supplying the data for the 1970 calculations.

The nature of the index is presented in Section I, illustrated by its application to 1966 tax data compiled by Joseph Pechman and Benjamin Okner. In addition, progressivity of 1966 taxes is compared with that of 1970 by comparing 1966 values of the index with those calculated from 1970 tax data compiled by Okner. Section II concludes with discussion of a number of properties of the index.

I. The Index of Tax Progressivity (S)

The nature of the index of tax progressivity is best illustrated by example. For this purpose, the data of Table 1 will be employed. Column 1 contains the percentage of families accumulated in order of income, marked off in deciles. Column 2 contains the corresponding accumulated percent of total income. The remaining columns contain the accumulated percent of total tax revenue contributed by these same families. For example, the third line of the table, corresponding to the 30 percent of families with the lowest incomes, indicates that these families received only 8.13 percent of total family income, contributed 2.90 percent of total revenue raised by the individual income tax, bore 4.38 percent of the corporate income tax, and so on.

The index of progressivity developed from these data is related to the Lorenz curve and the Gini concentration ratio. Although these measures are generally familiar, it will facilitate the presentation to begin with a short review of their nature. Data for the Lorenz curve of income distribution are contained in the first two columns of Table 1. When accumulated percent of total family income is plotted vertically against accumulated percent of families on the horizontal axis, we obtain the familiar Lorenz curve of Figure 1. It will be recalled that if income were exactly

TABLE 1—ACCUMULATED *U.S.* INCOME AND TAX BURDEN BY POPULATION DECILES, 1966

Population Decile	Cumulated Percentage: Adjusted Family Income	Individual Income Tax	Corporate Income Tax	Property Taxes	Sales and Excise Taxes	Payroll Taxes	Personal Property & Motor Vehicle Taxes	Total Federal Taxes	State and Local Taxes	Total Taxes
1	1.21	0.16	0.53	0.85	2.13	0.70	1.72	0.54	1.45	0.81
2	3.88	0.89	1.97	3.18	6.25	3.02	5.42	2.10	4.48	2.83
3	8.13	2.90	4.38	6.89	12.22	8.32	11.50	5.39	9.08	6.51
4	13.92	6.60	7.21	10.96	19.90	16.24	19.70	10.41	14.81	11.74
5	21.16	12.00	10.38	15.33	29.07	26.60	27.47	17.01	21.40	18.34
6	30.22	19.51	13.87	20.19	40.02	39.34	37.14	25.37	29.28	26.56
7	40.02	28.22	17.91	25.78	51.07	52.25	47.61	34.47	37.68	35.45
8	52.29	40.28	23.59	33.19	64.43	67.28	60.81	46.15	48.20	46.78
9	67.45	56.03	32.15	44.33	79.38	83.75	76.83	60.62	61.35	60.85
10	100	100	100	100	100	100	100	100	100	100
Addendum:										
Average tax rate		8.5	3.9	3.0	5.1	4.4	0.3	17.6	7.6	25.2

Source: Calculated from data in Pechman and Okner, and Okner. Adjusted family income is variant 1c Tax burden was calculated for each population decile by applying decile tax rates to adjusted family income for each decile. Results were then converted to percentages and accumulated.

equally distributed, the Lorenz curve would follow the diagonal line *OB*, but because the poorest 10 percent of families receive less than 10 percent of total income, the curve sags below the diagonal, following *OCB*. The greater the inequality of income, the farther the Lorenz curve bows away from the diagonal.

The Gini ratio measures income concentration by the proportion of the area of triangle *OAB* that is contained in the sector bounded by the diagonal line *OB* and the curve *OCB*. Thus the Gini ratio can vary between 0 for income equality to 1 for the extreme inequality in which all income is concentrated in a single family.

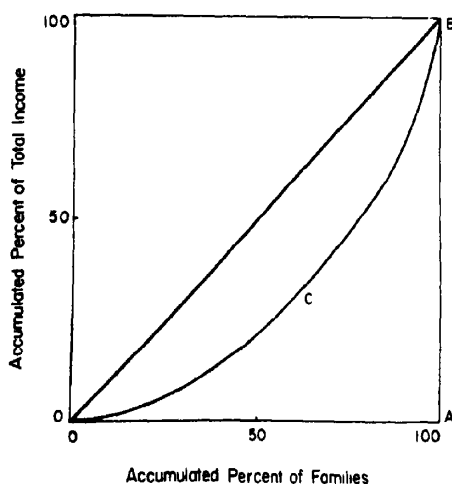



FIGURE 1. LORENZ CURVE OF *U S* FAMILY INCOME

To measure the progressivity of a tax, we employ a figure similar to a Lorenz curve, but one in which the accumulated percent of tax burden is plotted vertically against the accumulated percent of income on the horizontal axis. Such Lorenz curves are plotted in Figure 2 to represent the individual income tax and all sales and excise taxes combined. If the income tax were strictly proportional to income, the poorest 10 percent of all families, who earn 1.21 percent of all family income, would bear 1.21 percent of the income tax burden. The poorest 20 percent with 3.88 percent of family income would pay 3.88 percent of the tax, and so on. Thus the curve plotted for such a proportional tax would follow the diagonal line *OB*. Since, however, the income tax is progressive, the percentage of tax burden borne by the lowest income groups is smaller than their share of total income and the curve sags below the diagonal. Thus the curve *OCB* corresponds to the income tax.

In contrast, the percent of total tax burden imposed on low-income families by a

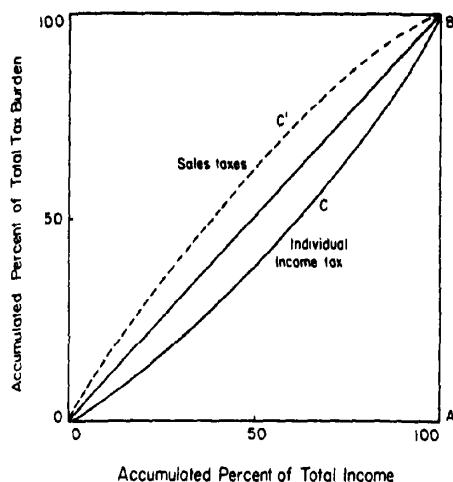



FIGURE 2. LORENZ CURVES FOR INDIVIDUAL INCOME TAX AND FOR ALL SALES AND EXCISE TAXES

regressive tax exceeds their percentage share of total income, so the curve $OC'B$, corresponding to the sales tax, arches above the diagonal.

Analogously to the Gini ratio we define the index of progressivity S in terms of K, the area of the triangle OAB, and L, the area $OABC$, contained between the Lorenz curve and the horizontal axis OA, so that

$$S = (K - L)/K = 1 - (L/K) \quad (1)$$

For a proportional tax, $L = K$, so $S = 0$. Since the curve corresponding to a progressive tax sags below the diagonal, the area L is smaller than K. As a result, the index S is positive for progressive taxes. In the limiting case where the highest income bracket bears the entire tax burden, the Lorenz curve lies along the sides OA and AB, so $L = 0$, and $S = +1$.

With a regressive tax, the Lorenz curve arches above the diagonal. This makes the area L larger than K, so S is negative. In the extreme case of regressivity, $L = 2K$ and $S = -1$. In other words, the index of progressivity S varies between $+1$ in the limiting case of progressive tax, through 0 for proportional taxes, to -1 in the limiting case of regressivity. Inspection of Figure 2 indicates that the values of S for income and for sales taxes are roughly equal in absolute magnitude, but with a positive index corresponding to the progressive income tax and a negative value corresponding to the regressive sales tax. This is borne out by measurement (see below) which yields for the income tax, $S = +.17$, but for the sales tax, $S = -.16$.

Figure 3 presents Lorenz curves for the six taxes studied by Pechman and Okner as of 1966. As would be expected from the known concentration of wealth as compared to income, taxes on corporate income and on property were much the most progressive *U.S.* taxes. Payroll taxes—mostly for social security and unemployment insurance—were the most regressive, although by only a small margin over sales taxes and taxes on personal property and motor vehicles. These findings are borne out by calculated indexes of progressivity as presented in Table 2. Progressivity of *U.S.* 1966 taxes varied from $S = +.36$ for the corporate income tax to $S = -.17$ for payroll taxes.

The second column of Table 2 shows corresponding values of S when calculated from *U.S.* tax data as of 1970 as compiled by Okner. Comparison of the two columns reveals interesting changes in the progressivity of individual taxes between the two dates. Whereas the individual income tax

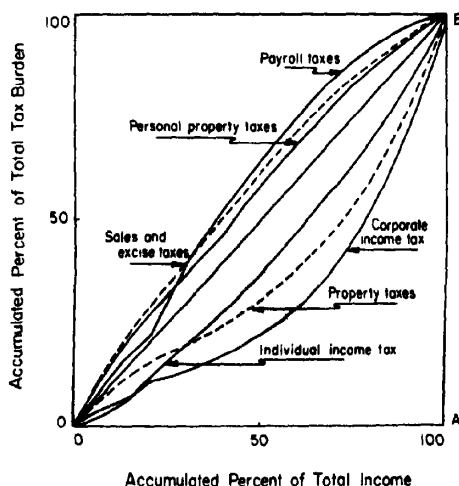



FIGURE 3. LORENZ CURVES FOR SIX *U.S.* TAXES

TABLE 2—PROGRESSIVITY OF U.S. TAXES, 1966 AND 1970

	Index (S) 1966	1970
Individual income tax	.17	.19
Corporate income tax	.36	.32
Property tax	.23	.18
Sales and excise taxes	-.16	-.15
Payroll taxes	-.17	-.13
Taxes on personal property and motor vehicles	-.12	-.09
All federal taxes	.087	.091
All state and local taxes	.045	.027
All taxes	.074	.070

became slightly more progressive, the corporate income tax became somewhat less progressive. Property taxes became considerably less progressive, while payroll taxes became considerably less regressive. Sales and excise taxes and taxes on personal property became less regressive.

Taken as a whole, the system of federal taxes became slightly more progressive over the period, but the state and local tax system became slightly less progressive. The two changes were nearly exactly offsetting, however, and the *U.S.* tax system as a whole retained in 1970 the same slight progressivity it had had in 1966.

II. Properties of the Index

A. *Mathematical Representation*

It facilitates exposition to represent the accumulated percent income, measured on the horizontal axis, as a variable y that ranges from 0 to 100. The ordinate of the Lorenz curve representing the corresponding accumulated percent of total tax burden for a given tax x, then becomes $T_x(y)$. In these terms, the area under the curve corresponding to tax x is given by

$$L_x = \int_0^{100} T_x(y)\,dy \tag{2}$$

Recalling that the area of triangle QAB has been designated K, we see that the index of progressivity of tax x is given by

$$\begin{aligned} S_x &= 1 - (L_x/K) \\ &= 1 - (1/K)\int_0^{100} T_x(y)\,dy \end{aligned} \tag{3}$$

B. *Calculation of the Index*

In practice, of course, the values of $T_x(y)$ are known for only a few discrete values of y. In Table 1, indeed, values are given for only 11 values of y: for $y_1, y_2, \ldots, y_{10}$, corresponding to the population deciles, and for $y_0 = 0$. But this information is adequate to provide a close approximation to the value of the integral as:

$$\begin{aligned} L_x &= \int_0^{100} T_x(y)\,dy \\ &\approx \sum_{i=1}^{10} (1/2)[T_x(y_i) + T_x(y_{i-1})](y_i - y_{i-1}) \end{aligned} \tag{4}$$

This approximation is easily calculated from data like that of Table 1, and has been employed in the compilation of Table 2. The area of the triangle, K, is of course the same for all taxes. Since the triangle has base and altitude of 100, $K = 5{,}000$ throughout.

C. *Response to Transfer of Tax Burden Among Families*

The index has the following important property: any change in tax law that transfers part of the tax borne by any family to another family with higher income increases S. Likewise any transfer of tax burden to a lower income family reduces S. Since this property of the index is quite obvious, it is sufficient to demonstrate it informally. Suppose a change in law transfers p percent of total tax burden from families in the third population decile of the income distribution to families in the seventh decile. Such a transfer leaves unaffected the values of $T_x(y_i)$ associated with deciles 1 and 2 and with deciles 7 through 10, but subtracts p from values of $T_x(y_i)$ associated with deciles 3 through 6. As can be seen from the

approximation (4), the result is a reduction of L and an increase in S. Similarly, transfer of tax burden from high income to low income families increases L and reduces S.

D. *Systems of Taxes*

Although the degree of progressivity of individual taxes is often of interest, it is more important from a policy point of view to treat the tax system as a whole. Another useful property of the index of progressivity is that the index for a system of two or more taxes is the weighted average of the indexes for the individual taxes, with respective average tax rates as weights. Average tax rate is defined for this purpose as the ratio of total dollar revenue yield of the tax to total dollar family income.

In other words, where S_x and S_z are indexes for taxes x and z, and r_x and r_z are the respective average tax rates, the index of progressivity of the two taxes taken together as a tax system, S_{xz}, is given as

$$(5) \quad S_{xz} = (r_x S_x + r_z S_z)/(r_x + r_z)$$

To see why this is so, recall that the ordinate of the Lorenz curve for tax x, that is, $T_x(y)$, is the accumulated percentage of total revenue from tax x borne by those families whose incomes accumulate to y percent of total income. If total dollar income for all families combined is Y, then total revenue from the tax in question is $r_x Y = R_x$. In these terms, the total burden of tax x borne by those families whose incomes accumulate to y percent of total becomes

$$(6) \quad \$_x = R_x T_x(y)/100 = r_x Y T_x(y)/100$$

Similarly, the total dollar burden of tax z on these same families is

$$(7) \quad \$_z = r_z Y T_z(y)/100$$

It immediately follows that the percent of total burden of the two taxes combined that is borne by these families is

$$(8) \quad T_{xz}(y) = (\$_x + \$_z)/(R_x + R_z) = [r_x T_x(y) + r_z T_z(y)]/(r_x + r_z)$$

It follows that

$$(9) \quad S_{xz} = 1 - (1/K) \int_0^{100} T_{xz}(y)\,dy = 1 - (1/K) \left[r_x \int_0^{100} T_x(y)\,dy + r_z \int_0^{100} T_z(y)\,dy \right] /(r_x + r_z)$$

It requires only slight algebraic manipulation to complete the proof.

When the individual tax indexes in Table 2 are averaged, using the average tax rates provided in the addendum to Table 1, the results for the entire tax system were $S = .077$ in 1966 and $S = .069$ in 1970, as close to the value calculated directly for the total tax system as rounding permits. When the entire *U.S.* tax system is treated as a whole, highly progressive taxes average with regressive taxes to produce a system that is very nearly proportional.

Since the subsystem of federal taxes is more heavily weighted with highly progressive taxes than is the state and local system, it proves to be the more progressive of the two, but the difference is surprisingly small, as can be seen when the two are plotted together in Figure 4. The values of S are

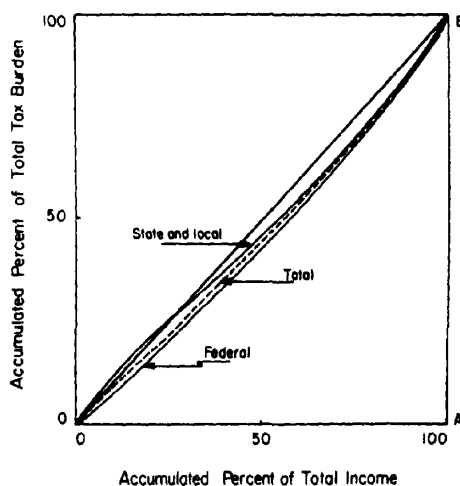



FIGURE 4. CURVES FOR FEDERAL, STATE AND LOCAL, AND TOTAL TAXES

recorded in Table 2. Note again that the value of S for the total tax system is the weighted average of the values for the federal and state and local subsystems.

E. *Distribution of Income and the Value of S*

In interpreting the index of progressivity, it must be borne in mind that the value obtained for any given tax is affected by the initial distribution of income among families. One simple way to see that this must be so is to consider a poll tax. The percent of poll tax burden borne by families with accumulated percent of income y is approximately proportional to the percent of families in that income bracket. Thus the Lorenz curve represenging the poll tax would be identical to the Lorenz curve of Figure 1, but with the axes interchanged. This makes the index S for the poll tax simply the negative of the Gini ratio for the income distribution. The poll tax represents an extreme instance, but similar considerations apply to any other tax.

This aspect of the index serves to emphasize that income distribution is central to the very concept of progressivity. There is nothing inherently regressive about a sales tax or even a poll tax. They are regressive because income is unequally distributed, and the more unequally income is distributed, the more regressive they become. Comparison of indexes for different taxes properly reflects the nature of these taxes in terms of the income distribution of the society within which they are applied. By the same token, however, the dependence of the index on income distribution presents an important qualification to its use in comparison of tax progressivity among different societies with different income distributions.

F. *The Index as an Average*

The index S measures the average progressivity of a tax or tax system across the entire income range, yet some taxes are progressive over one range of incomes and regressive over another. Figure 5 compares the curves of two taxes, one that is regressive at its lower but progressive at its upper range, the other is the opposite. The S ratio for such taxes depends on the algebraic sum of the areas between the curve and the diagonal line, with areas below the diagonal taken with positive, and those above the diagonal with negative sign. Thus both the taxes in Figure 5 would be described by $S = 0$, and would be classified as proportional taxes.

FIGURE 5 LORENZ CURVES OF TWO TAXES WITH $S = 0$

This is a familiar problem with any sort of average. Distributions with equal means can have widely different variances; those with equal variances can have widely different skewness. The S index is no exception to the general rule that it is impossible to capture completely a complex phenomenon by a single measurement, but it still represents a widely useful measure of tax progressivity when carefully applied.

REFERENCES

B. A. Okner, "Total U.S. Tax Burdens: 1966 and 1970 Compared," paper presented at the DFG Symposium, Augsburg, Germany, July 1976.

Joseph A. Pechman and Benjamin A. Okner, *Who Bears the Tax Burden?*, Washington 1974.

On the Optimal Size of Underpriced Facilities

By RICHARD C. PORTER*

Concern over the optimal size of a facility whose use is unpriced or underpriced has finally appeared (see Gene Mumy and Steve Hanke), but under the assumption that the use of the excessively demanded facility is costlessly rationed by some random mechanism such that "each of the demanded consumption units... has an equal probability of being satisfied" (p. 714). While examples of lotteries exist, by far the most common rationing devices for underpriced facilities are queues and congestion in which the user pays in wasted time rather than (or as well as) money, and those with the lowest willingness to pay are excluded.[1] This brief note outlines the efficiency criteria for determining the optimal size of facilities where use is rationed by means of wasted time.

When consumers pay time rather than money, the demand curve—i.e., the willingness-to-pay-money curve—is no longer the correct measure of the marginal benefit of an increased flow of users. In this note, the appropriate marginal benefits are identified and the criteria for the efficient size of the facility are derived for rationing by queue and by congestion. Finally, it is shown that welfare is greater in an optimally sized, underpriced, congested facility than in an optimally sized, underpriced, queued-for facility; and the conditions are explored under which queueing and congestion yield greater welfare than an optimally sized, underpriced facility rationed by random entry (i.e., that discussed by Mumy and Hanke).

I. Assumptions

First of all, it must be noticed that this is a second best problem since zero pricing of any scarce resource is per se inefficient—the burden of queueing, or waiting time, is a deadweight loss, and congestion will be excessive in the absence of tolls because of the negative reciprocal externalities.[2] Where the government[3] has *chosen* to underprice because some goal other than efficiency is thereby served, there is a question whether efficient investment criteria are at all relevant. When we seek such criteria, we must be assuming that the application of a higher price would be desirable but is precluded by some irresistible noneconomic force.[4]

There are inevitably resource costs to the admission or exclusion of customers for a limited facility, and these will differ among various rationing mechanisms. This note proceeds on the traditional assumption that all such costs are zero. Moreover, we should recognize that while both queueing and congestion waste time, they do not waste it in the same way. It is often possible to do something else while waiting—"English housewives... enjoy a good wait" (Nichols, Smolensky, and Tideman, p. 312)—and congestion often implies discomfort and reduced quality of service as well as wasted time.[5] Here we assume that time is

*Professor of economics, University of Michigan.
I would like to thank Alan Deardorff for several helpful comments on an earlier version.

[1] Failure to recognize that wasted time is a rationing process is most flagrant in Seneca, where zero willingness to pay users are assumed to enter the facility even after congestion arises.

[2] On waiting time, see D. Nichols, E. Smolensky, and T. N. Tideman; on congestion, see among others, Frank Knight or M. Bruce Johnson.

[3] While temporary underpricing due to misestimates of demand are not rare in the private sector (see, for example, the discussion of professional football in Roger Noll, pp. 141 ff), the persistent underpricing we are concerned with appears essentially in the province of public decision making.

[4] We ignore here, through the use of partial equilibrium analysis, any problems evoked by inefficiency elsewhere in the economy.

[5] In principle, of course, decayed quality can always be translated into a monetary equivalent. For an

time, whether wasted in a queue or wasted through congestion.

The assumption that time is time and that people prefer not to waste it, at least through queues and congestion, means that it can be linearly translated into money for each user.[6] But we must go still further if we are to infer the willingness to pay money of a *group* of consumers from the amount of time wasted in a queue or through congestion. For example, let the lines labeled A and B in Figure 1 represent for each of two users (A and B) the offer curves (i.e., loci of indifference) between wasted time and money price for their first use of a facility. If a money price is set so as just to induce the entry of both A and B (say once each) without wait, the price will be P_0; user B derives no consumer's surplus, but A's willingness to pay money exceeds the price charged by an amount equal to the distance A_0P_0. If, instead, no money price is charged and an admission queue lengthens sufficiently to just induce the entry of both A and B (again, once each), the time wasted by each will be T_0; now A derives no surplus, but B would have been willing to wait B_0T_0 more time-units for entry; the money value of that time to B is equal to the distance OB_1. But we cannot cavalierly assume that OB_1 equals A_0P_0. The money value of the consumer's surplus will be different for the same flow of users depending on whether time or money is used to ration entry. This is a serious complication, both conceptually and empirically, and this brief note avoids it by assuming that the offer curves of A and B (and indeed, of all potential users) are parallel. Then the money value of the consumer's surplus is the same regardless of whether the rationing mechanism is a money price or wasted time—but it does require us to assume that all potential users have the same rate of tradeoff between money expenditure and wasted time.[7]

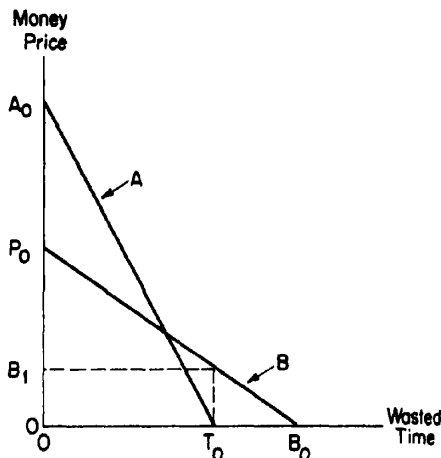



FIGURE 1

Of course, not all potential users need be identical in their evaluations of the facility. Indeed, if all demands were identical and if the facility were used to capacity before anyone used it a second time, there would be no consumer's surplus to consider regardless of whether the rationing were done by money or by time. What we assume (in terms of Figure 1) is that the price-time tradeoff lines of different users are parallel, but not coincident.

II. The Marginal Benefit Curve

The vertical axis of the demand curve for the facility can, under the above assumptions, be thought of as the willingness to pay in either money or time. Thus, in Figure 2, the flow of users (X) will be restrained to X_0 either by the imposition of a money charge of P_0 or by the development

interesting application of the differential subjective cost of waiting and congestion, see J. Vernon Henderson.

[6]For a rational consumer who chooses his hours of work, the linear translation can be made through the multiplication of time by either the after-tax wage rate or the marginal rate of substitution between leisure and income—since the two are identical. Otherwise, we must use a weighted average of the two rates. Further adjustment is needed when there is special disutility attached to wasted time or decayed quality. The linear relation assumes that the facility demands a sufficiently small part of the budget that income effects and inframarginal changes can be ignored.

[7]And, by choice of units, the rate of tradeoff is made to be one to one.

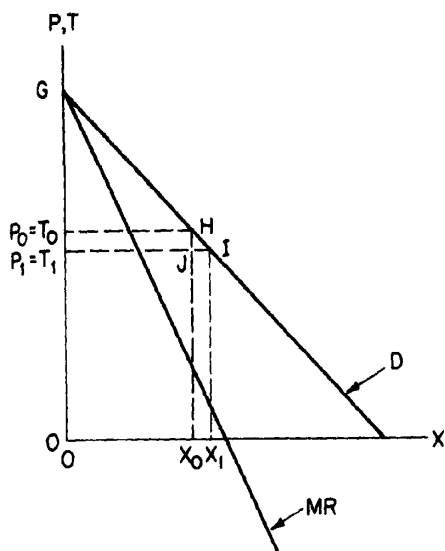



FIGURE 2

of a queue which requires wasted time of T_0 Consumers are indifferent whether a money price of P_0 or a time price of T_0 is used to ration the facility. In either situation, the same X_0 users enter the facility and the inframarginal users (and uses) receive the same aggregate consumer's surplus—equal to the triangle, P_0GH (or T_0GH).

But the similarity between a money price of P_0 and a time "price" of T_0 ends there. For when the money price is charged, someone collects the revenue of OP_0HX_0; the money value of the time wasted in the queue—also OP_0HX_0—accrues to no one. Thus, the surplus generated when a money price is charged is $OGHX_0$, but when a time price is "charged" is only P_0GH. In both cases, the consumer's share of the surplus is P_0GH, but what is the seller's gain when money is charged becomes a deadweight loss if a queue is formed.

The marginal benefit of admitting one more consumer to the facility is therefore not the same when entry is rationed by wasted time rather than by money. When money is the rationing mechanism, an increase in the flow of users from X_0 to X_1 requires a reduction in the money price from P_0 to P_1 and occasions an increase in the surplus generated of X_0HIX_1. The consumer's share of this additional surplus is P_0HIP_1, and the seller's share is X_0JIX_1 minus P_0HJP_1 (which is positive if marginal revenue is positive). All this is well known, but it is necessary to reproduce it to make clear the great difference when wasted time is the rationing mechanism. In order to induce an increased use of the facility from X_0 to X_1, a reduction in the wasted time is necessary, from T_0 to T_1. The inframarginal consumers gain, as with a reduction in the money price from P_0 to P_1, by an amount equal to P_0HIP_1. But no revenue accrues to anyone, only a change in the total time wasted in the facility.[8] Thus the change in the surplus generated is equal to the change in the consumer's surplus. When it is rationed by wasted time, the marginal benefit of expanding the use of the facility from X_0 to X_1 is T_0HIT_1—quite different from the marginal benefit of X_0HIX_1 when it is rationed by money.

In general, the marginal benefit of expanding the flow of users into a facility rationed by wasted time is the vertical distance between the demand curve and its concomitant marginal revenue curve. To show this, let the willingness to pay (P) be $P = P(X)$, where P' is the first derivative of P with respect to X, and the marginal revenue (MR) is $(XP' + P)$. The marginal benefit (MB_X) of entry under a rationing scheme of wasted time can then be derived:

$$\begin{aligned} MB_X &= \frac{d}{dX}\left[\int_{x=0}^{X} P(x)\,dx - P(X)\cdot X\right] \\ &= -XP' \\ &= P - MR \end{aligned}$$

Under money rationing, the marginal benefit of the facility is P, the money price the marginal user is willing to pay; under time rationing, the marginal benefit is $(P -$

[8]The time wasted per use is reduced from T_0 to T_1, but the number of uses is increased from X_0 to X_1. Whether the total time wasted rises or falls depends upon whether demand is elastic or inelastic.

MR), the time saved by the inframarginal users.

Very little can be said generally about the position and slope of this marginal benefit (MB_X) curve except that it will be below the demand curve when demand is price elastic and above the demand curve when demand is price inelastic.[9] Most disturbingly from the viewpoint of application, even the direction of slope of the MB_X curve is uncertain. The functional form of the demand curve—or more precisely, the exact relation of price elasticity to quantity—is critical. For example, the MB_X curves always slope downward (and are convex to the origin) for constant price elasticity demand curves, but they always slope upward (linearly from the origin) for linear demand curves.[10]

III. Rationing by Queue

Rationing by queue means in its clearest sense that the government (or some public spirited group) provides a certain size of the facility and then admits applicants with no (or small) money charge, but only as rapidly as is consistent with the absence of congestion in the facility. The queue of applicants lengthens until, in equilibrium, the waiting time discourages just enough applicants that the facility can handle the remaining flow without congestion. The size of the facility (S), defined as the flow of users it can accomodate without congestion, is therefore equal to the actual flow (X), and one can directly compare the benefits of admitting X to the costs of providing S in order to determine the optimal size of the facility.

A taxonomy of cases—i.e., of all combinations of upward-sloped and downward-sloped curves of marginal benefit of flow of users (MB_X) and marginal cost of size of facility (MC_S)—is not necessary. Three basic kinds of results can emerge; they are

FIGURE 3

illustrated in Figure 3(a, b, and c).[11] Figure 3a shows the case of an interior optimum where S^* ($= X^*$) lies between zero and one.[12] Figures 3b and 3c also display intersections of MB_X and MC_S, but at minima rather than maxima. Thus, the optima are at the extremes; in Figure 3b, no facility should be constructed (i.e., $S^* = X^* = 0$), and in Figure 3c, a facility adequate to handle the full zero price demand is called for (i.e., $S^* = X^* = 1$).

IV. Rationing by Congestion

The analysis is more complex for rationing by congestion because the actual flow

[9]It can be shown that $MB_X = P/\eta$, where η is the absolute value of the price elasticity of demand.

[10]The MB_X curve is even horizontal for one family of demand curves, that in which $X = k_1 e^{-k_2 P}$, where k_1 and k_2 are positive parameters. Then MB_X is equal to $1/k_2$ for all X.

[11]By suitable choice of flow units, the flow demanded at zero price is hereafter set equal to unity. It should be noted that consideration of demand curves which never reach the quantity axis, and hence imply an infinite demand at zero price, is precluded, but they are usually inappropriate for this kind of analysis.

[12]The asterisk hereafter indicates the optimum.

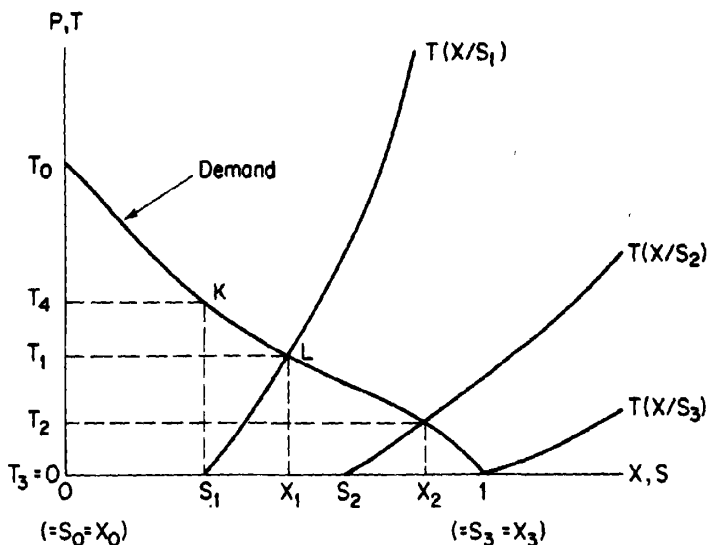



FIGURE 4

of users (X) does not necessarily equal the size of the facility (S). Indeed, since we have defined the "size" of the facility as the largest flow possible without congestion, rationing by congestion implies $X > S$. And where X exceeds S, it is not appropriate—as when analyzing queues—to compare straightforwardly the marginal benefit of an addition to X with the marginal cost of an identical addition to S. It is necessary to inquire into the way in which X changes in response to changes in S.

Let us make the assumption, partly for simplicity and partly for realism, that there are constant returns to scale in the facility; for example, a doubling of the size of the facility permits a doubling of the flow of users at any given degree of congestion. Then we can write time wasted through congestion as $T = T(X/S)$, where positive values of T' and T'' for all $X/S > 1$ reflect the ever greater average wasted time that accompanies greater congestion. In Figure 4 the effect of increasing S from zero through two intermediate values to one is illustrated.[13] Clearly, $X = S$ when S equals zero or one; in between, X will be greater than S. For any demand curve whose marginal revenue curve slopes downward, dX/dS will be greater than one for low values of S and less than one for high values of S.[14] Since the marginal cost of flow

[13] Recall that saturation (i.e., zero price) demand is defined as one. It should be noticed that the $T(X/S)$ lines represent average not marginal congestion cost and hence that the flows X_1 and X_2 indicate excessive congestion of facilities sized S_1 and S_2, respectively excessive, that is, from an efficiency viewpoint

[14] For a facility whose use is to be rationed by time wasted through congestion,

(a)	$P = T$	(by choice of units)
(b)	$P = P(X)$	(demand curve; $dP/dX = P' < 0$)
(c)	$T = T(X/S)$	(congestion time, assumed dependent on the ratio of X to S; $dT/d(X/S) = T' > 0$ and $d^2T/d(X/S)^2 = T'' > 0$)

In equilibrium,

(d) $$P(X) = T(X/S)$$

and hence

(e) $$\frac{dX}{dS} = \frac{(X/S)T'}{T' - SP'}$$

In the limit as S goes to zero, dX/dS goes to X/S which is greater than one (since X is always greater than S except at $X = S = 0$ and $X = S = 1$). And in the limit as S goes to one, dX/dS goes to $T'/(T' - P')$ which is less than one. Furthermore, dX/dS falls monotonically from a value above one at $S = 0$ to a value below one at $S = 1$ if d^2X/dS^2 is negative, which it will be if (as a sufficient but not necessary condition) the marginal revenue curve attendant to $P(X)$ is downward-sloped (i.e., if $XP'' + 2P' < 0$). (I thank Phillip S. Kott for preventing an error here.)

(MC_X) is related to the marginal cost of size (MC_S) by $MC_X = MC_S/(dX/dS)$, this means that MC_X will be less than MC_S for small S and greater than MC_S for large S (where "small" and "large" mean relatively near zero and one, respectively). Thus the MC_X curve will be more steeply upward-sloped, or less steeply downward-sloped, than the relevant MC_S curve.

Although the situation is more complicated with congestion than with a queue from an analytical viewpoint, the qualitative implications of the queue analysis are unchanged. Three kinds of results can emerge from the criteria for optimal size:

a) $0 < S^* < X^* < 1$

b) $X^* = S^* = 0$

c) $X^* = S^* = 1$

just as illustrated in Figure 3. But the fact that the MC_X curve is more upward-sloped (or less downward-sloped) than the MC_S curve does make it more plausible that situation a) will generally emerge, rather than b) or c).

V. Toward Empirical Application

The criteria for optimal size developed above can be straightforwardly applied—in principle. One must learn: 1) the demand curve, which permits calculation of the marginal benefit of use of the facility (MB_X); 2) the marginal cost of the size of the facility (MC_S); 3) the money value of wasted time to users of the facility; and 4) if congestion is to ration the flow, the way in which use reacts to size (dX/dS). One then applies the well-known marginal and total criteria.

While the principle is straightforward, the empirical task is not. Compared to money-priced facilities, a more careful estimate of the demand curve is called for since its functional form is so critical to estimating marginal benefits. Moreover, if congestion is contemplated as the rationing scheme, the complexities of exactly how use (X) will respond to size (S) must be explored. Finally, the money value of wasted time is neither the same for all users—as we have been assuming throughout—nor simple to estimate.[15] When different users value time differently, the relationship between the demand and the marginal benefit curves is even more intricate.

VI. Welfare Implications

It is well known that the first best rationing mechanism for any facility which is congestible[16] involves both congestion and a money price. Here we explore the second best welfare optima where a *single* rationing scheme is to be adopted and where the facility is of optimal size for that rationing scheme. Three non-money-price rationing methods are compared: queueing, congestion, and random entry. This last device, analyzed extensively by Mumy and Hanke, gives an equal likelihood of free admittance to each potential demander and hence implies a marginal benefit curve that is horizontal at the average willingness to pay of all potential demanders.

First, compare queueing with random entry. The marginal cost of entry under the two schemes is the same under both, since no congestion is permitted to arise (i.e., $X = S$). Which yields greater welfare[17] therefore depends entirely on the differences in marginal benefits. The comparison can be made by examining the four possible cases with an upward-sloped marginal cost of size of facility (MC_S) curve and a monotonic marginal benefit of flow of users under queue rationing (MB_X^Q) curve, as shown in Figure 5. The four cases differ as to whether the MB_X^Q curve is sloped upward (a and b) or downward (c and d) and whether the optimum size occurs to the left (a and c) or right (b and d) of the intersection of the two marginal benefit curves (MB_X^Q for queueing and MB_X^R for random entry). Nothing can be said in general about whether the optimal size will be larger

[15] See, for example, Jan Acton.

[16] By congestible is here meant simply that X can exceed S though at a cost of greater time per customer.

[17] That is, the excess of the total benefit of flow of users over the total cost of the facility.

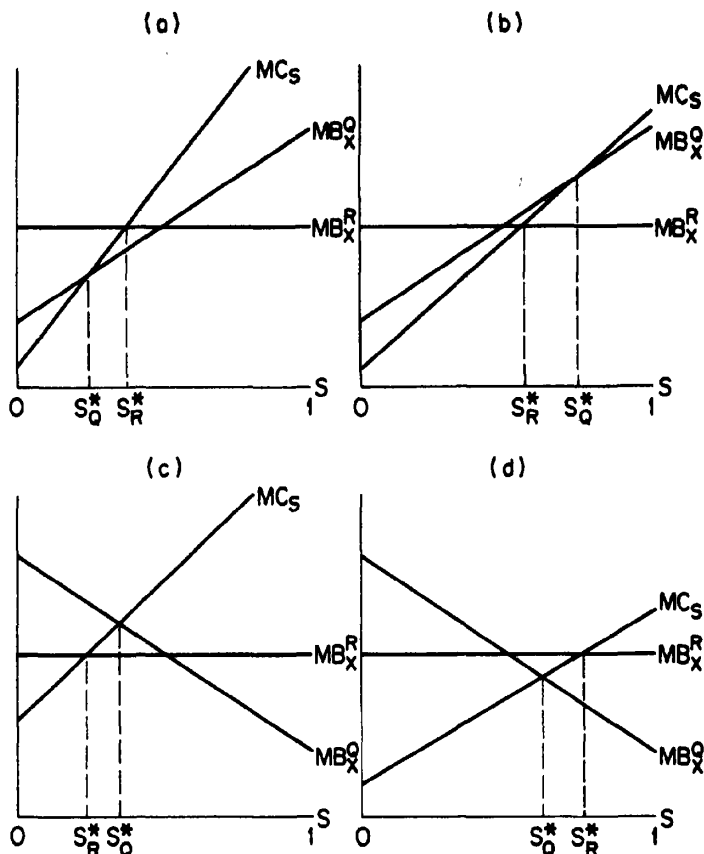



FIGURE 5

under queueing or random entry. But examination of the areas between the relevant MB_X curve and the MC_S curve (up to the optimal size) does show that total welfare will be greater with a queue if the MB_X^Q curve is sloped downward (c and d), and total welfare will be greater with random entry if the MB_X^Q curve is sloped upward (a and b).[18]

[18]Comparison of the areas of surplus is straightforward for Figures 5a and 5c; for the other two comparisons, it is necessary to recall that the total potential benefit is the same under any rationing mechanism—i.e., when all demanders are admitted free of any time or money cost, the consumer's surplus is equal to the total area under the demand curve. Downward-sloped MC_S curves do not alter the generalizations of the text except that they make corner solutions more likely. Obviously, if S_Q^* and S_R^* are both equal to 0 or 1, there is no difference in the welfare they generate.

Next, compare congestion with queueing. The marginal benefit of entry under the two schemes is the same under both, since wasted time is the rationing device. Which is superior therefore depends entirely on the differences in marginal cost.[19] The welfare comparison is readily made with the help of Figure 4. For *any* size of facility, other than S equal to zero or one, the average time wasted per user is greater if a queue is used to restrain the flow to S than if a congested larger flow is permitted; for example, in a facility of size S_1 (in Figure 4), the average time wasted is T_4 for a queue and only T_1 for congestion. Moreover, congestion permits a greater flow of users (i.e., $X_1 > S_1$). Thus, at S_1, the consumer's surplus under

[19]Recall that, for congestion, MC_X^C is *not* the same as MC_S^C (whereas MC_X^Q and MC_S^Q are the same).

congestion is larger than that under queueing by the area, T_4KLT_1. Differently put, a congested facility of size S_1 provides as much consumer's surplus as a queued-for facility of the larger—and hence more costly—size X_1. While this greater efficiency of congestion in generating surplus makes it impossible to say generally whether the optimally sized facility with congestion (S_C^*) is larger or smaller than that with queueing (S_Q^*), it is clear that at optimal size the welfare provided by congestion (W^C at S_C^*) is greater than the welfare provided by an optimally sized queued-for facility (W^Q at S_Q^*). The proof: W^C at S_C^* is greater than W^C at S_Q^* by the definition of optimal size; and W^C at S_Q^* is greater than W^Q at S_Q^* by virtue of the fact that, for any size of facility, more consumer's surplus is generated if it is congested than if it is queued for; therefore, W^C at S_C^* is greater than W^Q at S_Q^*.

Finally, compare congestion with random entry. The relative optimal sizes cannot be determined in general (i.e., $S_C^* \gtrless S_R^*$), as the preceding two paragraphs suggest. If the marginal benefit curve with congestion is downward-sloped, the maximum welfare with congestion exceeds that with random entry (since W^C at $S_C^* > W^Q$ at $S_Q^* > W^R$ at S_R^*). If the MB_X^C curve is upward-sloped, no general result emerges.

Thus, if an optimally sized facility is provided for the rationing mechanism used, congestion always offers greater welfare than queueing, and both congestion and queueing offer greater welfare than random entry if the marginal benefit curve (of queueing and congestion) is downward-sloped. Where this marginal benefit curve is upward-sloped, random entry yields greater welfare than queueing and *perhaps* greater welfare than congestion.

REFERENCES

Jan P. Acton, *Demand for Health Care Among the Urban Poor, with Special Emphasis on the Role of Time*, New York, April 1973.

J. V. Henderson, "Road Congestion: A Reconsideration of Pricing Theory," *J. Urban Econ.*, July 1974, *1*, 346–65.

M. B. Johnson, "On the Economics of Road Congestion," *Econometrica*, Jan./Apr 1964, *32*, 137–50.

F. H. Knight, "Some Fallacies in the Interpretation of Social Cost," *Quart. J Econ.*, Aug. 1924, *38*, 582–606; reprinted in George J. Stigler and Kenneth E Boulding, eds., *Readings in Price Theory*, Chicago 1952.

G. E. Mumy and S. H. Hanke, "Public Investment Criteria for Underpriced Public Products," *Amer. Econ. Rev.*, Sept. 1975, *65*, 712–20.

D. Nichols, E. Smolensky, and T. N. Tideman, "Discrimination by Waiting Time in Merit Goods," *Amer. Econ. Rev.*, June 1971, *61*, 312–23.

Roger G. Noll, "Attendance and Price Setting," in his *Government and the Sports Business*, Washington 1974, Ch. 4.

J. J. Seneca, "The Welfare Effects of Zero Pricing of Public Goods," *Publ. Choice*, Spring 1970, *8*, 101–10.

Welfare-Maximizing Price and Output with Stochastic Demand: Note

By PER ANDERSEN*

In a comment to Gardner Brown, Jr. and M. Bruce Johnson (hereafter called B-J), Michael Visscher points out that the B-J conclusions change if the rationing system changes. In the B-J approach it is assumed that available production is allocated to those with the highest consumer's surplus. Visscher analyzes two alternatives. In the first system service is offered first to those claimants with the least willingness to pay given a limited production. In the second system it is assumed that the available production is allocated randomly between all customers willing to pay the price $\bar{P}$. The purpose of this note is primarily to point out some unnoticed implications of Visscher's two rationing systems, but also to correct a minor error in his analysis. The notation is the same as that of Visscher.

I. A Reformulation

If service is first given to those customers with the lowest willingness to pay, the sum of consumer's surplus and total revenue can be depicted as the area $BCDE$ in Figure 1 in case of excess demand.[1] The objective function is assumed to be $W = E$ (willingness to pay $-E$ (average variable costs $\cdot$ sales) $-E$ (capacity costs). If the stochastic term of the demand function u is sufficiently high, the demand cannot be satisfied. Therefore L_1 and L_2 must be subtracted from the area $ACDO$ to obtain an expression for total willingness to pay. The expected value of $L_1 + L_2$ is

$$(1)\quad E(L_1 + L_2) = \int_{Z-X(\bar{P})}^{\infty} f(u) \left\{ \int_{X^{-1}(X(\bar{P})-Z)}^{X^{-1}(-u)} [X(P) + u]\, dP + [X(\bar{P}) + u - Z] X^{-1}(X(\bar{P}) - Z) \right\} du$$

The value of G in Figure 1 is determined from the relation

$$(2)\quad X(P) + u = X(\bar{P}) + u - Z$$

which gives

$$(3)\quad G = X^{-1}(X(\bar{P}) - Z)$$

and not, as indicated by Visscher,

$$(3')\quad G = X^{-1}(-u) - X^{-1}(Z - u) + P$$

The welfare expression is given in (4).

The objective function is maximized by differentiating with respect to $\bar{P}$ and Z and by setting the derivatives equal to zero as done in equations (5) and (6).

$$(4)\quad W = \int_{-\infty}^{\infty} f(u) \int_{\bar{P}}^{X^{-1}(-u)} [X(P) + u] \cdot dP\, du + \bar{P}X(\bar{P}) - \int_{Z-X(\bar{P})}^{\infty} f(u) \int_{X^{-1}(X(\bar{P})-Z)}^{X^{-1}(-u)} [X(P) + u]\, dP\, du - \int_{Z-X(\bar{P})}^{\infty} f(u) X^{-1}(X(\bar{P}) - Z) \cdot [X(\bar{P}) + u - Z] \cdot du - b\left\{ X(\bar{P}) - \int_{Z-X(\bar{P})}^{\infty} u f(u) du + [Z - X(\bar{P})] \int_{Z-X(\bar{P})}^{\infty} f(u) du \right\} - \beta Z$$

$$(5)\quad \frac{\partial W}{\partial \bar{P}} = X'(\bar{P})[\bar{P} - (1 - F(Z - X(\bar{P}))) \cdot X^{-1}(X(\bar{P}) - Z) - bF(Z - X(\bar{P}))] = 0$$

*Assistant professor, University of Odense, Denmark. I am indebted to E. Gørtz and S. Holm for valuable comments on a first draft of this paper.

[1] It is assumed that every potential customer demands only one unit of the product.

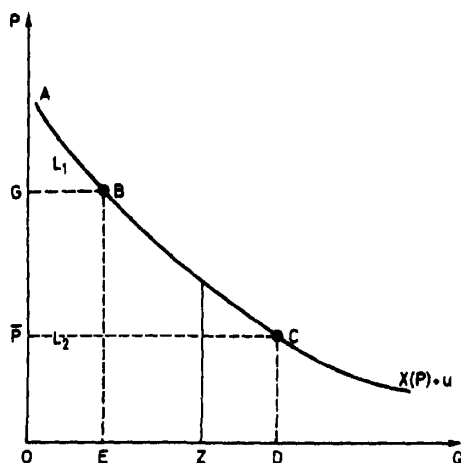



FIGURE 1

$$(6)\quad \frac{\partial W}{\partial Z} = (1 - F(Z - X(\bar{P}))) \cdot (X^{-1}(X(\bar{P}) - Z) - b) - \beta = 0$$

$$(7)\quad \left.\frac{\partial W}{\partial Z}\right|_{\substack{\bar{P}=b+\beta \\ Z=X(b+\beta)}} = (1 - F(0)) \cdot (X^{-1}(0) - b) - \beta$$

Adding (5) divided by $X'(\bar{P})$ to (6) we obtain that the optimal price given an optimally adjusted capacity becomes $\bar{P} = b + \beta$. Depending on cost and demand parameters the optimal capacity may be smaller or greater than riskless capacity. This can be seen by evaluating $\partial W/\partial Z$ for $\bar{P} = b + \beta$ and $Z = X(b + \beta)$. From equation (7) it follows that the optimal capacity is greater (less) than the riskless capacity if the probability of excess demand multiplied by the difference between the riskless demand curve's intersection with the ordinate axis and the short-run marginal costs is greater (less) than the marginal capacity costs. In the special case where u is distributed symmetrically around zero, i.e., $F(0) = 1/2$, the optimal capacity becomes greater than in the riskless case if

$$(8)\qquad X^{-1}(0) > 2\beta + b$$

II. The Price-Capacity Relation

In Visscher's analysis it is assumed that capacity is optimally adjusted. The same holds for the B-J approach. But as the capacity adjustment is typically a time-consuming process, it is also of importance to analyze (5) given that Z is not necessarily optimally adjusted. It follows directly from (5) that the optimal price is always greater than the short-run marginal costs. It is of special interest to note that the price is not necessarily a decreasing function of the available capacity as is the case in normal economic models. As an illustration of this point it is assumed that the stochastic term u takes on two alternative values, u_1 and u_2 ($u_1 > u_2$), with the probabilities s and $1 - s$, respectively, and that the riskless demand is linear ($X(P) = A - BP$). From (5) we get that the optimal price P_1^* is given by equation (9). It follows from (9) that the optimal price is an increasing function of the available capacity within certain limits when those customers with the lowest willingness to pay are served first.

In my earlier paper, I show that this conclusion is valid also in the case where customers are served randomly. Under the same assumptions as above the optimal price P_2^* is shown in (10).

$$(9)\quad P_1^* = \begin{cases} \dfrac{A + u_2 - Z}{B} \text{ for } \dfrac{s}{(1-s)}\dfrac{Z}{B} + b < \dfrac{A + u_2 - Z}{B} \\ \dfrac{s}{(1-s)}\dfrac{Z}{B} + b \text{ for } \dfrac{A + u_2 - Z}{B} \leq \dfrac{s}{(1-s)}\dfrac{Z}{B} + b \leq \dfrac{A + u_1 - Z}{B} \\ \dfrac{A + u_1 - Z}{B} \text{ for } \dfrac{s}{(1-s)}\dfrac{Z}{B} + b > \dfrac{A + u_1 - Z}{B} \end{cases}$$

$$(10)\quad P_2^* = \begin{cases} \dfrac{sZ}{2(1-s)B} + b \text{ for} \\ \dfrac{A + u_2 - Z}{B} \leq \dfrac{sZ}{2(1-s)B} \\ + b \leq \dfrac{A + u_1 - Z}{B} \\ P_1^* \quad \text{otherwise} \end{cases}$$

An economic interpretation which can be applied to both of the models considered above is given in my 1974 article.

In the B-J model it is shown that the optimal price is equal to short-run marginal costs. Given the special cases outlined above it is demonstrated that the optimal price within certain limits, in case of random rationing, is equal to the average of the optimal prices obtained in case of the two other refusal systems.

REFERENCES

P. Andersen, "Public Utility Pricing in Case of Oscillating Demand," *Swed. J. Econ.*, Dec. 1974, *76*, 402–14.

G. Brown, Jr. and M. B. Johnson, "Public Utility Pricing and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119–28.

M. Visscher, "Welfare-Maximizing Price and Output with Stochastic Demand: Comment," *Amer. Econ. Rev.*, Mar. 1973, *63*, 224–29.

Equilibrium Concepts in the Theory of Public Goods

By BRUCE N. ANGIER AND THOMAS S. MCCALEB*

In a recent article in this *Review*, Theodore Bergstrom and Robert Goodman formulate a model of the demand for public goods which they use to estimate certain parameters of the demand function. They claim that the assumptions on which their model is based were proposed and studied by Howard Bowen, and refer to an allocation of goods satisfying these assumptions as a Bowen equilibrium. We shall argue that this is a misrepresentation of Bowen's model; that Bergstrom-Goodman's assumptions are, in fact, not equivalent to Bowen's; and that this difference in assumptions reflects very different objectives. Where Bergstrom-Goodman's model is used to provide a positive analysis of collective choice under simple majority rule with fixed tax shares, Bowen was interested in the essentially normative problem of attaining efficiency in collective choice under some of the more common voting rules.

Their model is based on the following assumptions:

1: The public good can be provided at constant unit cost.
2: Each individual's share in the cost of public goods supply is independent of the amount of public expenditures and of his own preference revelation.
3: Each individual is able to determine his preferred quantity of the public good given his tax share.
4: The quantity supplied of any public good is equal to the median of the quantities demanded by the individuals.
5: The median quantity is the quantity demanded by the individual with median income.

*Instructor in economics, North Carolina State University, and assistant professor of economics, Rice University and University of Kansas, respectively. Appreciation is expressed to James A Wilde for a helpful comment on an earlier version.

The thrust of these assumptions is that the median voter rule from the theory of simple majority voting can be used to determine the outcome of the collective decision process in which the quantity of a public good to be provided is at issue. Having listed their assumptions, Bergstrom-Goodman go on to state that assumptions 1-4 were also proposed and studied by Bowen. Furthermore, they refer to an allocation satisfying assumptions 1-4 as a Bowen equilibrium.

Bowen's model is based on the following four assumptions (p. 34):

1: "...all individuals in the community actually vote and...each expresses a preference which is appropriate to his individual interests."
2: "...the cost to the community of providing various possible quantities of [public goods] is known."
3: "...the cost of whatever amount...is to be 'produced' will be divided equally among all the citizens."
4: "...the several curves of individual marginal substitution are distributed according to the normal law of error."

Bowen's assumptions 1 and 2 are similar, though not identical, to Bergstrom-Goodman's assumptions 1-3. Nowhere in Bergstrom-Goodman's analysis, however, does one find any analog to Bowen's third and fourth assumptions. Moreover, Bowen's analysis is not confined to simple majority voting outcomes, as is the Bergstrom-Goodman discussion. Instead, Bowen takes into account both simple majority voting and plurality voting.

Assumptions 3 and 4 of the Bowen model are sufficient to ensure that the collectively chosen output of public goods will be that quantity for which the summed marginal rates of substitution just equal

marginal cost under either of the above mentioned voting rules. This is, of course, the efficient level of provision—the "ideal output" in Bowen's terms. Thus, Bowen's equilibrium consists of a distribution of cost shares among individuals such that, given the voting rule and the distribution of individual preferences, the community's choice of a public goods output will be efficient according to the Pareto criterion. Since Bowen believed, on the basis of some limited evidence available to him at the time, that individual demands for public goods in fact approximated a normal distribution, this suggests that his objective was to examine conditions under which the tax shares were so distributed as to ensure efficiency in collective decisions regarding the provision of public goods under the more common voting rules.

Bergstrom-Goodman's concept of equilibrium in contrast is nothing more than the result of applying the median voter rule. No assumptions are made about the distribution of preferences, nor are any conditions placed on the distribution of tax shares. Hence, the collectively chosen output is unrestricted. In particular, there is nothing in the Bergstrom-Goodman notion of equilibrium to generate efficiency in collective decisions regarding the quantity of public goods. Bergstrom-Goodman's objective of investigating the outcome of a simple majority voting rule under a given arbitrary distribution of tax shares seems very different from Bowen's objective of investigating conditions on the distribution of tax shares which lead to efficiency in the voting outcome.

This difference in objectives, embodied in the differing assumptions of the two models, gives rise to an important dichotomy. Bergstrom-Goodman's concept of a Bowen equilibrium is, as they note, not in general Pareto optimal. By contrast, Bowen's own solution does satisfy the Pareto criterion. In the context of Musgrave's analytical framework, Bowen's model is efficient, but it need not be equitable since it does not require that individual tax shares be equal to individual marginal evaluations of the public goods output level. In fact, under his assumptions the distribution of tax shares is necessarily inequitable. Bergstrom-Goodman's concept, on the other hand, is in general neither efficient nor equitable. Moreover, it is a positive rather than a normative solution.

Bergstrom-Goodman's confusion regarding the relationship between their analysis and Bowen's model also leads them to suggest the Lindahl equilibrium as a normative criterion against which their outcome may be measured in terms of efficiency. The Lindahl equilibrium, however, is unnecessarily restrictive for this purpose as it requires not only that the output of the public good be efficient, but also that the tax shares be distributed among individuals in an equitable manner. If only efficiency in public goods supply is at issue, as it seems to be with Bergstrom-Goodman, the appropriate *conceptual* comparison would be between the output level chosen in their model and the output level which is chosen in Bowen's model. For this reason it appears useful to maintain the distinction between the median voter equilibrium on which Bergstrom-Goodman's attention is focused, and the efficient though not necessarily equitable equilibrium of the model examined by Bowen.

REFERENCES

T. C. Bergstrom and R. P. Goodman, "Private Demands for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280–96.

H. Bowen, "The Interpretation of Voting in the Allocation of Economic Resources," *Quart. J. Econ.*, Nov. 1943, *58*, 27–48; reprinted in Kenneth Arrow and Tibor Scitovsky, eds., *Readings in Welfare Economics*, Homewood 1969.

Richard A. Musgrave, *The Theory of Public Finance*, New York 1959.

Multiperiod Consumption-Investment Decisions: Further Comments

By WILLIAM T. ZIEMBA*

In a correction of his 1970 paper, Eugene Fama (1976) has shown that existence of some value of β_{t+1}, having positive probability, for which his inequality (4) holds strictly is sufficient to verify that the derived utility function U_t inherits the strictly concave property of U_{t+1}. This corrects the difficulty referred to in my 1974 note. In the deterministic case, for fixed β_{t+1}, the function $U_{t+1}(C_t, H_t R(\beta_{t+1})' \mid \beta_{t+1})$ will not generally be strictly concave in the vector H_t when U_{t+1} is strictly concave in C_t and $w_t \equiv H_t R(\beta_{t+1})'$ since the function $H_t R(\beta_{t+1})'$ is not 1:1 (when there are more than two securities). However, when one takes the expected value of U_{t+1} with respect to β_{t+1}, the integral of U_{t+1} will be strictly concave in H_t if $H_t R(\beta_{t+1})'$ is 1:1 with positive probability. One may clarify what this means by considering the case when there are three states of nature and three securities. Then the difficulty occurs when[1]

$$(1) \quad \begin{aligned} H_1^* R_1^1 + H_2^* R_2^1 + H_3^* R_3^1 &= \tilde{H}_1^* R_1^1 + \tilde{H}_2^* R_2^1 + \tilde{H}_3^* R_3^1 \\ H_1^* R_1^2 + H_2^* R_2^2 + H_3^* R_3^2 &= \tilde{H}_1^* R_1^2 + \tilde{H}_2^* R_2^2 + \tilde{H}_3^* R_3^2 \\ H_1^* R_1^3 + H_2^* R_2^3 + H_3^* R_3^3 &= \tilde{H}_1^* R_1^3 + \tilde{H}_2^* R_2^3 + \tilde{H}_3^* R_3^3 \end{aligned}$$

has a solution $(H^*, \tilde{H})$ and $H^* \neq \tilde{H}$. It is clear that for (1) to only have the solution $H^* - \tilde{H} = 0$, the matrix

$$R \equiv \begin{pmatrix} R_1^1 & R_2^1 & R_3^1 \\ R_1^2 & R_2^2 & R_3^2 \\ R_1^3 & R_2^3 & R_3^3 \end{pmatrix}$$

must have rank equal to 3 since (1) may be written as

$$R(H^* - \tilde{H})' = 0$$

Thus if two securities have equal or proportional returns in all states of nature, then the strict concavity results fails. In general, for discrete distributions having m possible occurrences, if there are n securities to guarantee strict concavity it is necessary and sufficient that there be exactly n linearly independent return vectors each having positive probability of occurrence, i.e., the $m \times n$ matrix R has rank n. For general distributions one must assume that if $Pr[H_t^* R(\beta_{t+1})' = \tilde{H}_t R(\beta_{t+1})'] = 1$, then $H_t^* = \tilde{H}_t$. This is then equivalent to a full rank assumption; with say a joint normal distribution of returns it is tantamount to assuming that the variance-covariance matrix is positive definite. Certainly there are many possible return distributions for which the strict concavity reduction property will fail to hold.

As a historical note it should be pointed out that the concavity preserving reduction results discussed by Fama (1970,1976) have their origin in George Dantzig's work on stochastic programming and Richard Bellman's work on stochastic dynamic programming in the early 1950's. For a totally rigorous proof of the reduction results one would need to prove that the function U_t is well defined, i.e., measureable and that an optimum allocation vector H_t^* exists in each period t. These aspects are discussed by R. Tyrell Rockafellar and Roger J.-B. Wets, and by Hayne E. Leland.

*Faculty of Commerce and Business Administration, University of British Columbia, and Operations Research Unit, Marmara Scientific and Industrial Research Institute, Gebze, Turkey This research was supported by grant 66-0250 from the Canada Council.

[1]For simplicity I supress the time subscripts and let R_j^i refer to the ith possible realization of security j.

REFERENCES

Richard Bellman, *Dynamic Programming*, Princeton 1957.

G. B. Dantzig, "Linear Programming Unde

Uncertainty," *Manage. Sci.*, Jan. 1955, *3*, 197–206.

E. F. Fama, "Multiperiod Consumption-Investment Decisions," *Amer. Econ. Rev.*, Mar. 1970, *60*, 163–74.

———, "Multiperiod Consumption-Investment Decisions: A Correction," *Amer. Econ. Rev.*, Sept. 1976, *66*, 723–24.

H. E. Leland, "On the Existence of Optimal Policies Under Uncertainty," *J. Econ. Theory*, Feb. 1972, *4*, 35–44.

R. T. Rockafellar and R. J.-B. Wets, "Stochastic Convex Programming: Basic Duality," *Pac. J. Math.*, Jan. 1976, *62*, 173–95.

W. T. Ziemba, "The Behavior of a Firm Subject to Stochastic Regulatory Review: Comment," *Bell J. Econ.*, Autumn 1974, *5*, 710–12.

On the Theory of the Competitive Firm Under Price Uncertainty: Note

By YASUNORI ISHII*

Agnar Sandmo, in his excellent paper, attempted to analyze the behavior of the competitive firm under price uncertainty. He presumed that: (a) the utility function of the firm is a concave, continuous and differentiable function of profits, so that

$$(1) \qquad U'(\pi) > 0, \qquad U''(\pi) < 0$$

and the absolute risk aversion $R_A(\pi)$ $(= -U''(\pi)/U'(\pi))$ is a decreasing function of π, that is,

$$(2) \qquad \partial R_A(\pi)/\partial\pi < 0$$

(b) The firm's profit function can be defined as

$$\pi(x) = Px - c(x) - B$$

where P is the price of output, assumed to be a (subjectively) random variable with density function $f(P)$ and expected value $E[P] = \mu$, x is output, $c(x)$ is the variable cost function with $c'(x) > 0$ and $c(0) = 0$, and B is fixed cost.

(c) The expected utility of profits can be written as

$$E[U(Px - c(x) - B)]$$

where $E[\cdot]$ is the expectations operator. The competitive firm decides the output so as to maximize the expected utility of profits.

I

Sandmo analyzed the effects of change in the expected value and uncertainty of the price on the optimal output of the competitive firm, which were written as $\partial x/\partial\theta$ and $\partial x/\partial\gamma$ in his article, under the assumptions mentioned above. As a result of his analysis, he asserted that while the sign of $\partial x/\partial\theta$ can be clearly judged to be positive, the sign of $\partial x/\partial\gamma$ is ambiguous in general.

The purpose of this note is to show that investigating the sign of $\partial x/\partial\gamma$ under the same assumptions and analytic apparatus as Sandmo's, we can decide the sign of $\partial x/\partial\gamma$, manifestly in opposition to his assertions.

II

Necessary and sufficient conditions for a maximum are written as

$$(3) \qquad E[U'(\pi)(P - c'(x))] = 0$$

$$(4) \quad D = E[U''(\pi)(P - c'(x))^2 - U'(\pi)c''(x)] < 0$$

If $c''(x) \geqq 0$, $D < 0$ is always satisfied. And there is a possibility of $D < 0$ even in the case $c''(x) < 0$. It is interesting to examine the conditions of existence and uniqueness of interior optimal solution. We, however, assume for the simplification of analysis that (3) and (4) determine a nonzero, finite, and unique solution x^* to the maximization problem.

Following the introduction of lemmas, we shall investigate the sign of $\partial x^*/\partial\gamma$.

LEMMA 1:

$$\mu - c'(x^*) > 0$$

LEMMA 2:

$$E[(P - c'(x^*))U''(\pi^*)] \geqq 0$$

where[1] $\quad \pi^* = Px^* - c(x^*) - B$

The proofs of these lemmas are omitted here for they are shown in Sandmo's paper.

[1]We assume that $R_A(\pi)$ is a nonincreasing function of π.

*Yokohama City University, Japan.

III

Now that all of the necessaries have been presented, we shall proceed to prove our proposition.

PROPOSITION: *Nonincreasing absolute risk aversion is a sufficient condition for $\partial x^*/\partial\gamma$ to be negative.*

PROOF:

By adopting the same procedures that Sandmo used, we can obtain

$$\frac{\partial x^*}{\partial\gamma} = -\frac{1}{D}\{E[U'(\pi^*)(P-\mu)] + x^*E[(P-c'(x^*))U''(\pi^*)(P-\mu)]\} \quad (5)$$

which is equal to his equation (13). Since $D < 0$, the sign of $\partial x^*/\partial\gamma$ is equivalent to that of $E[U'(\pi^*)(P-\mu)] + x^*E[(P-c'(x^*))U''(\pi^*)(P-\mu)]$. Consequently we will investigate the sign of this.

First, modifying $E[U'(\pi^*)(P-\mu)]$, we obtain

$$\begin{aligned}(6)\quad E[U'(\pi^*)(P-\mu)] &= E[(U'(\pi^*) - E[U'(\pi^*)])(P-\mu)] \\ &= cov(P, U'(\pi^*))\end{aligned}$$

where $cov(P, U'(\pi^*))$ is the covariance between P and $U'(\pi^*)$. Differentiating $U'(\pi)$ with respect to P at $x = x^*$ and taking into consideration $U''(\pi) < 0$, we can obtain

$$\frac{\partial U'(\pi)}{\partial P} = x^*U''(\pi^*) < 0$$

which means

$$cov(P, U'(\pi^*)) < 0 \quad (7)$$

in the case when the competitive firm faces price uncertainty. We therefore have, substituting (7) for (6),[2]

$$E[U'(\pi^*)(P-\mu)] < 0 \quad (8)$$

Next, from $E[(P-c'(x^*))U''(\pi^*)(P-\mu)]$, we can get

$$\begin{aligned}(9)\quad E[(P-c'(x^*))U''(\pi^*)(P-\mu)] &= E[(P-c'(x^*))(P-c'(x^*) + c'(x^*) - \mu)U''(\pi^*)] \\ &= E[(P-c'(x^*))^2\cdot U''(\pi^*)] \\ &\quad - (\mu - c'(x^*))E[(P-c'(x^*))U''(\pi^*)]\end{aligned}$$

It is clear immediately that the first term of the right-hand side of (9) is negative and that the second term of the right-hand side of (9) is nonpositive in consideration of Lemma 1 and Lemma 2. We therefore can obtain

$$x^*E[(P-c'(x^*))U''(\pi^*)\cdot(P-\mu)] < 0 \quad (10)$$

It follows that substituting (8) and (10) in (5) and considering $D < 0$ from (4), we have $\partial x^*/\partial\gamma < 0$.

[2]We can show another proof of (8)

$$\begin{aligned}E[U'(\pi^*)(P-\mu)] &= E[U'(\pi^*)(P - c'(x^*) + c'(x^*) - \mu)] = E[U'(\pi^*)(P - c'(x^*))] \\ &\quad - (\mu - c'(x^*))E[U'(\pi^*)] \\ &= -(\mu - c'(x^*))E[U'(\pi^*)] < 0\end{aligned}$$

REFERENCES

Kenneth J. Arrow, "The Theory of Risk Aversion" in his *Essays in The Theory of Risk-Bearing*, Chicago 1971.

J. W. Pratt, "Risk Aversion in The Small and in The Large," *Econometrica*, Jan./Apr. 1974, *32*, 122–36.

A. Sandmo, "On the Theory of the Competitive Firm under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

Externalities in a Regulated Industry: The Aircraft Noise Problem

By JEROLD B. MUSKIN AND JOHN A. SORRENTINO, JR.*

Airline noise is an externality in the traditional sense of being a by-product of normal economic activity. It affects the population around airports in single event bundles and cumulatively. The transient nature of noise itself is unlike most other externalities. As with the others, however, general "improvements" in technology and increases in population have made the situation more difficult to tolerate. The combination of these things has led to a critical situation for the affected population since it involves its physical and mental health.

From among the various methods of dealing with externalities, we choose the effluent charge scheme. It generally allows each firm to incorporate the environmental standard into its marginal operating choices. If each firm makes efficient decisions, then the cost to society of achieving an environmental standard will be a minimum. In this paper, there are two modifications of the traditional charge scheme. One is that because of the well-known difficulties in specifying and estimating social costs, we use the (estimated) direct noise abatement costs of achieving the environmental standard. The second is that airlines cannot be necessarily thought of as cost minimizers.

The charges used for the control of airline noise are generated as shadow prices in a simple linear programming model. Solution to the problem involves choosing a mix of noise abating options that achieves proposed environmental standards at least cost. Additional bounds on the problem are limitations on service reductions and rate of return (*ROR*) on investment regulation. The data sources used for the program were generally fragmentary and incomplete for our purposes.

After calculating a noise charge for a hypothetical airport, we discuss the implementation and ramifications of the charge plan in the regulated airline industry. Included is a discussion of the link between noise abatement and fuel consumption.

I. The Language of Aircraft Noise

Without going into detail about how they are derived, we use two related measures of noise: 1) the *effective perceived noise level in decibels* (*EPNdB*) which for single noise events is a subjectively adjusted measure of noise determined by human reaction; and 2) the *noise exposure forecast* (*NEF*) which represents the cumulative noise level during a 24-hour period. Bolt, Beranek and Newman, Inc. have shown that for aircraft class i and flight path j:

$$(1) \quad NEF_{ij} = EPNdB_{ij} + 10\,log\,[d_{ij} + 16.67 n_{ij}] - 88$$

where d_{ij}, n_{ij} are the number of daytime and nighttime flights, and 88 is a normalization factor. *NEF* at a ground point is then the sum over aircraft classes and flight paths.

$$(2) \quad NEF = 10\,log \sum_i \sum_j antilog \frac{NEF_{ij}}{10}$$

Populations exposed to equal *NEF* levels define *NEF* contours around an airport. The near-term goal of the Environmental Protection Agency (*EPA*) is to relieve all populations within *NEF* 45 to avoid long-term hearing effects. Since the *EPA* has never established any precise population removal goals, we create some for our model

*Assistant professor of marketing/transportation, Drexel University, and assistant professor of economics, Temple University, respectively.

Editor's Note: This paper was presented to the Eighty-Ninth Annual Meeting of the American Economic Association, and was printed in the February 1977 *Proceedings* issue of this *Review* (pp. 347–50). Because of substantial errors in layout, the paper is reprinted here, with our apologies to the authors.

II. The *LP* Model

The choice variable in the linear programming model is the set of options for noise abatement. These are broken down into operational and retrofit. The operational options are: (X_1) a composite of reduced thrust takeoffs, power cutback departures, flap management approaches, higher altitude approaches, two-segment approaches, and thrust reverse limitations; (X_2) preferential runways; (X_3) preferential flight paths; (X_4) night curfews; (X_5) aircraft type limitations; (X_6) aircraft weight limitations; and (X_7) acquisition of land areas.

The retrofit options are: (X_8) acoustically treating the *JT3D* engine; (X_9) extending X_8 to the *JT8D* engine; (X_{10}) installing front fans and larger housings on the *JT8D* engine and (X_{11}) combination of X_8 and X_{10}.

The planning period was taken to be 1974-85, and all dollar magnitudes were discounted back to 1974 at both 10 percent and 20.5 percent discount rates. Let x_j be the percent level of operation of option X_j and c_j the discounted present value of the cost of operating x_j at the 1 percent level. The objective function is

$$C = \sum_{j=1}^{11} c_j x_j \tag{3}$$

Each option X_j has a specified potential reduction efficiency, a_{ij}. Let P_i be the original population and p_i the population permitted to remain in contour i. The term $P_i - p_i$ is the population to be removed. Once removed the population enters contour $i + 1$. The four contours i are *NEF* 45, 40, 35, and 30. The constraint may be specified as

$$\sum_{j=1}^{11} A_{i+1,j} x_j \geqq \bar{p}_{i+1}, \quad i = 0,1,2,3 \tag{4}$$

where $A_{i+1,j} = [P_{i+1} a_{i+1,j} - P_i a_{ij}]$ is the "net migration" of contour $i + 1$, and $\bar{p}_{i+1} = [P_{i+1} - p_{i+1}]$ is the allowed residual in contour $i + 1$.

Another constraint is a maximum rate of reduction in seat-miles flown. If s_j is the percent reduction in seat-miles, $\bar{s}$ the maximum permissible reduction, z_0 the initial number of seat-miles, then we have

$$\sum_{j=1}^{11} s_j x_j \leqq \bar{s}/z_0 \tag{5}$$

The rate-of-return (*ROR*) constraint, as defined by the Civil Aeronautics Board (*CAB*), applies to *investment*, is a *minimum* and applies to the *industry* as a whole. It is currently set at 12 percent. If $R(x)$ is revenue, $C(x)$ operating cost, and $I(x)$ investment given the choice of option bundle x, then the *ROR* constraint is

$$\frac{R(x) - C(x)}{I(x)} \geqq V \tag{6}$$

We estimated the effects of the adoption of noise abatement options on R, C, and I via rough measures of fare, cost, and investment elasticities.

Using *IBM*'s MPS *360* package and data gotten from the cited references and elsewhere, we solved the *LP* problem. The package contained a sensitivity apparatus for the percent of goal (*POG*) achieved, perturbations in the population removal goals themselves, changes in the maximum service reduction, and changes in a fuel price index. The solutions were displayed for various *POG* levels (60–75 percent), fuel price indices (100–400), and the 10 percent and 20.5 percent discount rates. The solution included the optimal option bundle and total abatement costs. For the median year 1979, we give an example *ROR* calculation using 12 percent. It was found, for example, that with a fuel price index of 100, a *POG* of 65 was achieved at least cost ($822 million at 10 percent) with options X_1 to X_5 used at 100 percent, and X_{10} used at 3.7 percent of full-option use.

Upon perturbing the population removal constraints by 1,000 persons, we obtained shadow prices to be used as noise charges. Since the amount of noise is expressed in 10 *log* equivalent terms, we expressed the charge in dollars per these units.

The disaggregation of aircraft and flight path segments by airlines allows the contri-

bution of each airline to total noise to be determined and controlled by the charge. This extends equation (2) by summing the NEF_{ijk} over carriers k.

III. Policy Implications

Since the data were available on a national scale but charges must be imposed at each airport, we create a hypothetical three-carrier airport, which is represented by 5 percent of the national program. At the 10 percent discount rate and a fuel price index of 100, we calculated a charge of 36¢ per 10 *log* equivalent unit. The 36¢ charge is assessed on each airline according to its monitored noise output beyond the allowed levels. The charge applied to the *NEF* 35 contour which was shown to be the "critical" contour in the *LP* program. The reception of the charge by the airlines themselves leads to a discussion of the structure of the industry.

The airline industry is a regulated oligopoly. Inefficiency is due not only to the market structure but to the solidification of this structure through regulation. The most pronounced symptoms of this inefficiency are the overcapacity problem, the excessive rate of equipment innovation, and the proliferation of flights at preferred departure times. Two primary causes for these strategies are the proscription of price competition and the possibility of the *CAB* approving the introduction of competitors on "inadequately serviced" routes.

Airline performance can be attributed in part to three well-known behavioral theories: Oliver Williamson's expense preference effect, Harvey Leibenstein's *X-inefficiency* syndrome and the Averch-Johnson (A-J) overcapitalization effect. Williamson's thesis is that management may have goals other than cost minimization. In the airline industry, evidence of this is the proclivity for the introduction of new generation aircraft and the tendency not to diversify business investments outside of the airline industry itself. The X-inefficiency phenomenon may be characterized in the industry by imperfections in managerial performance due to lack of knowledge, insufficient incentives, and/or sloth. The traditional A-J thesis must be modified with respect to the airline industry. The *ROR* is defined on investment rather than the capital stock. The *ROR* limit is a minimum rate for the industry. The usual A-J behavior exhibited by individual firms is obscured. If the situation exists that the industry *ROR* is near the 12 percent minimum and overinvestment causes the *ROR* to dip below it, then fare increases will generally be allowed. It appears that since there is no limit on how high the *ROR* can be in this case, the industry would not be induced to hover around the minimum simply to obtain price increases. There is, however, a more subtle bound on how high *ROR* may go. When the *ROR* is high, the *CAB* is generally moved to introduce competitive carriers on the relevant routes. Hence, overinvestment to lower the apparent *ROR* may prevent carriers from facing competition. Airlines appear to regard this as a significant impetus.

Expenditures on noise abatement may also become subject to expense preference, X-inefficiency, and the A-J effect. The effect of X-inefficiency would most likely be attributed to a lack of perceptiveness of firm and social needs. The other two are more difficult to separate. Airline management may have preferences for or against ex-expenditures on environmental protection for reasons other than profits. Overinvestment in noise abatement, including overadoption of new aircraft, can either cause fare increases (at minimum *ROR*) or forestall the introduction of competitive carriers according to modified A-J behavior. Overinvestment in abatement can enhance the image of the firm which may have pecuniary (A-J) and nonpecuniary (expense preference) benefits. They still remain, however, economically inefficient.

The introduction of aircraft noise abatement to the situation of regulated inefficiency has several dimensions. The close technological and economic regulation of the industry by the Federal Aviation Administration (*FAA*) and the *CAB* would

facilitate the activities of a Noise Abatement Authority (*NAA*). The *NAA* can use existing monitoring techniques and the clear demarcation of aircraft to attribute noise to particular airlines for charge purposes. There is also much common technological knowledge so that possible "reaction functions" to the charges might be estimated. The *NAA* must project current cost data for each year of the planning period. This will enable it to publish a tentative schedule of year-to-year charges, allowing the airlines to make long-range decisions regarding investment, scheduling, pilot training, etc.

The feedback of airline responses to the charge in terms of noise abatement affects the linear program by changing the original populations in each contour. With these new constraints, the program can be recalculated and new noise charges established. A problem with the airline responses is that we cannot generally expect cost-minimizing behavior. Hence, along with the charge proposal, we also require that the *CAB* change regulatory policies by allowing greater competition to induce airlines (at least) to minimize costs.

Society's most general goal is to maximize the overall well-being of its citizens. We do not have a precise social welfare index with all of society's variables including airline service and airline noise as elements. In lieu of placing an explicit value on the benefit or detriment society receives from these, we simply fix their levels and find the least-cost way of achieving them. One particularly useful aspect of the analytical-computational model used above is that the "meta-problem" can be dealt with through perturbations of the parameters in the "inner problem." Although still given in terms of costs, society's broader tradeoff decisions can be explicitly seen.

The effluent charge scheme proposed in this paper can be an efficient operational procedure if accompanied by changes in regulatory procedures by the *CAB* inducing cost minimization. The cooperation of federal and state agencies, the airline industry, and local communities will increase the convergence toward a social optimum. The reflection in the exposed populations of changes in airline noise output make the *LP* problem sensitive to airline abatement activities. This allows efficient planning for capital-intensive operations. A noise abatement strategy that fails to take account of the cumulative nature of aircraft operations, the multiplicity of noise-abating options, the diversity of the social effects of noise (and its reduction), and the efficiency of the charge scheme is bound to yield a high-cost partially effective outcome. The current approach of the *EPA* is an example.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

W. J. Baumol, "On Taxation and the Control of Externalities," *Amer. Econ. Rev.*, June 1972, *62*, 307–21.

——— **and A. K. Klevorick**, "Input Choices and Rate of Return Regulation," *Bell J. Econ.*, Autumn 1970, *1*, 162–90.

Richard E. Caves, *Air Transport and Its Regulators: An Industry Study*, Cambridge, Mass. 1962.

W. E. Fruhan, Jr., *The Fight for Competitive Advantage: A Study of the Domestic Trunk Carriers*, Cambridge, Mass. 1972.

W. E. Jordan, *Airline Regulation in America: Effects and Imperfections*, Baltimore 1970.

A. V. Kneese and B. T. Bower, "Standards, Charges, and Equity," in *Managing Water Quality: Economics, Technology, Institutions*, Baltimore 1968.

H. Leibenstein, "Organization of Frictional Equilibria, X-Efficiency and the Rate of Innovation," *Quart. J. Econ.*, Nov. 1969, *82*, 600–23.

Oliver E. Williamson, *The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm*, Englewood Cliffs 1964.

Bolt, Beranek and Newman, Inc., *Aircraft Noise*

Analysis for the Existing Air Carrier System, rept. no. 2218, project no. 118992, Report to the Aviation Advisory Committee, Washington, D.C., September, 1972.

IBM Mathematical Programming System/360, Version 2, Linear and Separable Programming—User's Manual, Program no. 360A-CO-14X, 3d ed., White Plains, July 1971.

Serendipity, Inc. Airport/Aircraft System Noise, *A Study of the Magnitude of Transportation Noise Generation and Potential Abatement*, Vol. III, prepared for Department of Transportation, Office of Noise Abatement, rept. no. OST-ONA-71-1, Washington, Nov. 1970.

U.S. Civil Aeronautics Board, Domestic Passenger Fare Investigation, Testimony of J. Maldutis, Jr., Exhibit TW-T-B, Phase 7, Docket 21866-7, 1971.

U.S. Environmental Protection Agency, "Regulation of Aircraft Noise," Proj. Rept, Washington, July 1974.

ERRATA

Sons of Immigrants: Are They at an Earnings Disadvantage?

By BARRY R. CHISWICK

American Economic Review, Papers and Proceedings,

February 1977

Page 378, Table 2.

The regression coefficients or *t*-ratios for several key variables were not printed in the last two regressions:

TABLE 2—ANALYSIS OF EARNINGS OF NATIVE BORN WHITE MALES, 25 TO 64 YEARS OF AGE, BY NATIVITY OF PARENTS, 1970

	Parents Native or Foreign Born	
	(6)	(7)
PARFOR	0.04866 (4.67)	(a)
FAFOR	(a)	0.07688 (4.12)
MOPFOR	(a)	0.03735 (1.49)

Note (a) Variable not entered; *t*-ratios in parentheses

REPORT OF THE TREASURER FOR THE YEAR ENDING DECEMBER 31, 1976*

The financial position of the American Economic Association is the best it has been since the end of 1969. The surplus for 1976 was $150,000 with an even larger surplus in prospect for 1977. The net worth of the Association at more than $300,000 exceeds one-third of annual expenditures. (See Tables 1 and 2 and the financial statements, particularly the Statement of Assets and Liabilities, following the Auditors' Report.) Rising costs, however, can be expected to wipe out the surplus in a few years even if no new programs of expenditure are undertaken.

The surplus of $150,000 in 1976 was partly the result of two special items. The Association's claim against Richard D. Irwin, Inc. was settled, resulting in a payment to the Association of royalties on the Readings Series of $42,000. Publication of three volumes of the *Index of Economic Articles* resulted in a net gain shown in Table 1 of $38,000. (The Statement of Revenues and Expenses following the Auditors' Report shows gross sales of $51,000 under revenues and prorated printing costs of $13,000 under expenses.) This figure, however, needs explanation. It does not mean that there was a large profit—or any profit—on the *Index*. The Association spent $59,000 in 1976 for printing the three volumes. Only $13,000 is shown in the Statement of Revenues and Expenses as a cost, the remainder being considered an investment in inventories of unsold copies. Sales in future years will presumably enable the Association to recover the remainder of the printing costs and more (especially since we received payment in 1976 from the distributor of the *Index* only for copies sold through October 31, which did not include any sales of Volume XI). We hope that eventually the sales of the *Index* will return part of the editorial costs incurred in the office of the *Journal of Economic Literature*. We do not, however, expect to recover all the costs of the *Index*. Essentially what is happening in 1976–77 is a return of costs partly incurred in prior years. Since the pace of publication of the *Index* will eventually slow down to one volume a year, the large contribution of the *Index* to the surpluses of 1976–77 is not likely to be maintained.

Expenses for 1976 include provision of $39,000 for federal income tax. The actual liability may turn out to be smaller, but the auditors thought it prudent to make provision for the possibility that the payment by Richard D. Irwin, Inc., in settlement of the Association's claim, and the receipts from sales of the *Index* would be regarded as "unrelated business income" and therefore taxable. The provision for federal taxes in the budget for 1977 is considerably smaller because nothing comparable to the settlement with Richard D. Irwin, Inc. is in prospect.

The Association has been accumulating a reserve for a new *Directory* (or Handbook). The reserve is now $100,000. In view of the improved financial position of the Association, a new *Directory* is being planned for 1978. Its publication will have little effect on the 1978 surplus or deficit, but it will reduce the net worth of the Association. Consequently the net worth at the end of 1978 is likely to be little greater than at the end of 1976. In view of the relentless increase in costs, the present favorable financial position of the Association will be temporary unless expenditures are restrained or dues increased.

At its meeting on March 18, 1977, the Executive Committee approved the budget

**Editor's Note:* The reports of the Treasurer, the Finance Committee, and the Auditors of the American Economic Association were prepared too late for inclusion in the February 1977 *Proceedings* of this *Review*.

for 1977 shown in Table 1. (Its implications for cash flow are shown in Table 2.) In view of the prospect of rising costs and the rather low level of net worth relative to annual expenditures, the Executive Committee voted to increase dues and subscriptions 5 percent effective January 1, 1978.

RENDIGS FELS, *Treasurer*

TABLE 1—AMERICAN ECONOMIC ASSOCIATION BUDGET, ACCOUNTING BASIS, FOR 1977
(Thousands of dollars)

	1977 Budget	1976 (Actual)[a]	1975 (Actual)[a]
REVENUE			
Operating Income			
Dues and subscriptions	750	726	617
Advertising	80	76	64
Sales—miscellaneous	30	28	26
Sales—mailing list	35	35	33
Annual meeting	0	14	7
JOE	21	21	17
Sales of *Index* (net)	60	38[b]	-
Other income	41	41	23
Subtotal.	1,017	979	787
Settlement of claim against publisher	-	42	-
Investment income			
Interest and dividends	35	33	33
Real capital gains (losses)	39	(37)	(81)
Subtotal:	74	(4)	(48)
TOTAL REVENUE	1,091	1,017	739
EXPENSES			
Publications			
American Economic Review	231	211	196
JEL and *Index*	295	283	253
Papers and Proceedings	59	59	51
Handbook (*Directory*)[c]	50	47	61
JOE	25	25	22
Subtotal	660	625	583
Operating and Administrative			
Salaries	108	98	104
Rent	8	8	8
Committees	39	29	23
Other	96	106[d]	78
Subtotal:	251	241	212
TOTAL EXPENSES	911	866	795
SURPLUS (DEFICIT)	180	150	(56)

[a]The amounts in the columns for actual results for 1975 and 1976 are identical with those in the Statement of Revenues and Expenses accompanying the auditors' report, but some of the line items have been rearranged and consolidated.

[b]Sales receipts ($51,000) net of prorated share ($13,000) of printing costs ($59,000).

[c]Each year $50,000 is budgeted for a future *Directory*.

[d]Includes provision of $39,000 for federal income tax.

TABLE 2—AMERICAN ECONOMIC ASSOCIATION CASH BUDGET, 1977
(Thousands of dollars)

	1977 Budget	1976 Actual
SURPLUS (DEFICIT), Accrual basis	180	150
Plus noncash charges to accrual budget		
Reserve for *Directory*	50	50
Capital losses (gains)	(39)	37
Depreciation	1	1
Subtotal, surplus (deficit) adjusted to cash basis	192	239
OTHER OPERATIONS AFFECTING CASH		
Increase (decrease) in deferred income	20	41
Increase (decrease) in accounts payable, etc.	0	16
Decrease (increase) in accounts receivable	118	(117)
Decrease (increase) in prepaid expenses	-	(1)
Cash receipts less disbursements from restricted funds[a]	(5)	(30)
Subtotal, other operations affecting cash	133	(91)
TOTAL CHANGES FROM OPERATIONS	325	148
INVESTMENT-TYPE TRANSACTIONS		
Decrease (increase) in inventory of *Index* volumes	(30)	(46)
Sales (purchases) of office equipment	(1)	(1)
Changes in investment accounts:		
(Interest and dividends)	(35)	(33)
Transfers of cash from (to) investment accounts	(262)	(145)
Custodian and investment counsel fees	3	3
TOTAL INVESTMENT-TYPE TRANSACTIONS	(325)	(222)
INCREASE (DECREASE) IN CASH	0	(74)

[a]Excludes Economics Institute.

REPORT OF THE FINANCE COMMITTEE*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 31, 1976, with costs and market values as of that date. The total market value of the securities portfolio at year-end was $491,201. After making adjustments for cash additions and withdrawals (including a sizeable withdrawal of $60,000 made from the portion of the Fund represented by the Ford Foundation grant), we estimate that the Association's investment portfolio (on a total return basis) increased in value by 13.5 percent during 1976.

It should be remembered that the $491,201 figure referred to above includes a Special Grant that was made by the Ford Foundation in January 1969 and subsequently commingled with the Association's account. As of December 31, 1976, the Association's portion of the aggregate account was $438,234, or 89.2 percent and the Special Grant represented the remaining $52,967, or 10.8 percent of the total.

As was indicated in last year's report, the Finance Committee established a policy of including a combination of both common stocks and commercial paper in the investment portfolio rather than its previous all-equity orientation. The first moves in the direction of this new policy were made during 1976, bringing the year-end equity ratio to 70.8 percent. In addition, as mentioned above, a $60,000 withdrawal occurred during the year. As a result of these two factors, significant sales were made in the following issues: IBM, CBS, Abbott Laboratories, Eastman Kodak, First Bank Systems, Minnesota Mining, Moore Corporation, and Utah International. Additionally, in December, Utah International was acquired by General Electric.

In terms of the portfolio's appreciation during 1976, the 13.5 percent increase referred to above, while significant, did trail the widely followed market indices. This result was due in part to the more conservative equity ratio adopted by the Committee a year ago and also to the fact that the high quality growth stocks held in the portfolio underperformed the general stock market over this period despite continued strong increases in both the profits and dividends of these companies. However, including this most recent experience, it is interesting to note that, during the past ten years, the total return on the equities held in the Association's portfolio exceeded the comparable return of both the Dow Jones Industrial Average and the Standard and Poor's 500.

As we move into 1977, the Finance Committee is anticipating another year of favorable economic conditions characterized by moderate, but sustained, economic growth. This continued economic strength will be reflected in increased corporate profits and corporate dividend payments. Moreover, at the current level of the market, common stocks do not appear to be overvalued on the basis of underlying earnings, cash flow, and book value. Therefore, the Committee believes that a meaningful exposure to equities in the Association's investment portfolio continues to be warranted.

To be more specific, a year ago the Finance Committee decided to shift the portfolio over time to a 50 percent equity, 50 percent fixed-income orientation with a 66 percent equity ratio to be achieved by the end of 1976. At its November 23, 1976, meeting the Committee modified that decision somewhat and agreed to maintain the equity proportion at approximately two-thirds while permitting the flexibility to vary the ratio 5 percentage points in either direction. In view of the shifting economic trends as well as the beginning of a new administration in Washington, the Finance Committee will be monitoring developments closely and will make appropriate adjustments in investment policy as changes in the investment environment occur.

BERYL W. SPRINKEL, *Chairman*

*The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some discrepancies between figures in the Report of the Finance Committee and the Auditor's Report which follows.

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 31, 1976

	Value	Percent	Estimated Income
Cash Equivalents and Short-Term Securities	$143,435	29.2	$ 7,166
Medium-Term Securities	0	0.0	0
Long-Term Securities and Preferred Stocks	0	0.0	0
Convertible Securities	0	0.0	0
Equity Securities	347,766	70.8	11,092
Total	491,201	100.0	18,258

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 31, 1976

	Amount	Price	Value	Unit Cost	Total Cost	Est. Income
CASH EQUIVALENTS AND SHORT-TERM SECURITIES (29.2 percent)						
Cash Equivalents (0-1 Year) (29.2 percent)						
Cash			$117		$117	
Stein Roe Cash Reserves, Inc.	143,318	1	143,318	1	143,318*	$7,166
Subtotal Cash Equivalents			143,435		143,435	7,166
Total Cash and Fixed Income Securities			143,435		143,435	7,166
EQUITY SECURITIES (70.8 percent)						
Utilities (5.8 percent)						
Central and Southwest	1,200	17	20,250	10	11,720*	1,440
Banks (11.1 percent)						
Citicorp	500	33	16,375	29	14,531*	480
First Bank System	500	44	22,125	19	9,295*	760
			38,500		23,826	1,240
Other Financial (5.5 percent)						
Alexander and Alexander	500	38	19,063	19	7,325*	530
Foods and Containers (11.7 percent)						
Philip Morris	400	62	24,700	44	17,726	520
Seven Up	500	32	15,875	28	14,125	700
			40,575		31,851	1,220
Mining and Metals (6.3 percent)						
MAPCO	500	44	22,000	18	8,855	450
Oil and Gas (13.1 percent)						
Continental Oil	600	38	22,500	16	9,868	720
Gulf Oil	800	29	23,100	17	13,321	1,440
			45,600		23,189	2,160
Drugs and Medical (11.5 percent)						
Abbott Lab	400	49	19,650	35	14,136	400
Merck	300	68	20,438	52	15,631*	450
			40,088		29,767	850
Electrical Products (6.2 percent)						
General Electric	390	56	21,694	25	9,707*	702
Computers (8.0 percent)						
IBM	100	279	27,914	80	7,985*	900
Miscellaneous (20.7 percent)						
CBS	300	59	17,813	29	8,570*	600
Disney	334	47	15,782	2	817	40
Eastman Kodak	250	86	21,500	84	20,907*	525
Minnesota Mining and Mfg.	300	57	16,987	61	18,400*	435
			72,082		48,694	1,600
TOTAL EQUITY SECURITIES	—		347,766		204,919	11,092
TOTAL SECURITIES AND CASH			491,201		348,354	18,258

*More than one cost basis.

AUDITORS' REPORT

To the Executive Committee of
The American Economic Association:

We have examined the statements of assets and liabilities of THE AMERICAN ECONOMIC ASSOCIATION (a District of Columbia corporation, not for profit) as of December 31, 1976 and 1975, and the related statements of revenues and expenses, changes in general and restricted fund balances, and changes in assets and liabilities for the years then ended. Our examination was made in accordance with generally accepted auditing standards, and accordingly included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the accompanying financial statements present fairly the assets and liabilities of The American Economic Association as of December 31, 1976 and 1975, and its revenues and expenses, changes in fund balances and the changes in its assets and liabilities for the years then ended, in conformity with generally accepted accounting principles consistently applied during the periods.

Arthur Andersen & Co.

Nashville, Tennessee
February 22, 1977

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF ASSETS AND LIABILITIES—DECEMBER 31, 1976 AND 1975

Assets	1976	1975
CASH	**$ 46,373**	**$121,950**
INVESTMENTS, at market (Notes 1 and 2):		
Temporary investments	405,763	245,000
Permanent investments	491,084	486,731
	896,847	**731,731**
ACCOUNTS RECEIVABLE:		
Advertising, back issues, etc.	76,845	53,412
Sales of *Index of Economic Articles*	51,303	-
Receivable from publisher	41,924	
Allowance for doubtful accounts	(1,000)	(1,100)
	169,072	**52,312**
INVENTORY OF *Index of Economic Articles* (at cost)	**46,098**	-
PREPAID EXPENSES	**5,405**	**4,251**
OFFICE FURNITURE AND EQUIPMENT (at cost) less accumulated depreciation of $6,174 in 1976 and $5,144 in 1975	**7,027**	**7,413**
Total Assets	**$1,170,822**	**$917,657**

Liabilities and Fund Balances	1976	1975
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	**$ 168,015**	**$152,136**
DEFERRED INCOME (Note 1):		
Life membership dues	68,034	65,480
Other membership dues	282,852	255,992
Subscriptions	169,967	157,818
Job Openings for Economists	11,049	11,908
	531,902	**491,198**
ACCRUAL FOR *Directory* (Note 1)	**100,000**	**50,000**
FUND BALANCES:		
Restricted (Note 4)	48,595	139,676
Add (deduct)–Unrecognized change in market value of investments (Notes 1 and 3)	9,067	(7,456)
	57,662	132,220
General	277,198	110,535
Add (deduct) Unrecognized change in market value of investments (Notes 1 and 3)	36,045	(18,432)
General fund-net worth	313,243	92,103
Total fund balances	325,793	250,211
Add (deduct) Unrecognized change in market value of investments (Notes 1 and 3)	45,112	(25,888)
Net fund balance	**370,905**	**224,323**
Total Liabilities and Fund Balances	**$1,170,822**	**$917,657**

The accompanying notes to financial statements are an integral part of these statements

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1976 AND 1975

	1976	1975
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$ 468,444	$380,471
Nonmember subscriptions	257,273	237,191
Job Openings for Economists subscriptions	21,034	16,852
Advertising	76,032	63,786
Sale of *Index of Economic Articles*	51,303	-
Sale of copies, republications, and handbooks	27,664	25,964
Sale of mailing list	35,202	33,397
Annual meeting	29,064	12,824
Sundry	40,921	22,709
	1,006,937	793,194
SETTLEMENT OF CLAIM AGAINST PUBLISHER	41,924	-
INVESTMENT LOSSES (Note 2)	(4,281)	(48,125)
Net Revenues	**1,044,580**	**745,069**
PUBLICATION EXPENSES:		
American Economic Review	210,962	195,627
Journal of Economic Literature	283,012	253,407
Papers and Proceedings	58,518	50,989
Directory publication (Note 1)	47,455	60,903
Job Openings for Economists	24,926	21,788
Index of Economic Articles	12,807	-
	637,680	582,714
OPERATING AND ADMINISTRATIVE EXPENSES:		
General and administrative:		
Salaries	98,032	103,996
Rent	8,078	7,854
Other (Exhibit 1)	67,026	62,443
Committee	28,639	22,618
Annual meeting	15,398	8,915
Provision for federal income taxes (Note 6)	39,352	12,200
	256,525	218,026
Total Expenses	**894,205**	**800,740**
REVENUES IN EXCESS OF (LESS THAN) EXPENSES	**$ 150,375**	**$(55,670)**

The accompanying notes to financial statements and Exhibit 1 are an integral part of these statements.

The American Economic Association Statement of Changes in General Fund Balance for the Years Ended December 31, 1976 and 1975

	Total	Operations	Market Value Adjustments
Balance at January 1, 1975	**$148,381**	**$ (66,609)**	**$214,990**
Add—market value adjustments resulting from inflation (Note 1)	17,825	-	17,825
Deduct—expenses in excess of revenues	(55,671)	(55,671)	-
Balance at December 31, 1975	**110,535**	**(122,280)**	**232,815**
Add—market value adjustments resulting from inflation (Note 1)	16,288	-	16,288
Add—revenues in excess of expenses	150,375	150,375	-
Balance at December 31, 1976	**$277,198**	**$ 28,095**	**$249,103**

The accompanying notes to financial statements are an integral part of this statement.

The American Economic Association Statement of Changes in Restricted Fund Balances for the Year Ended December 31, 1976

	Balance at January 1	Receipts	Disbursements	Allocation of Investment (Losses) (Note 4)	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics	$102,004	$ 2,367	$ (62,367)	$(1,260)	$40,744
The Alfred P. Sloan Foundation, Chase Manhattan Bank, and Ford Foundation grants for increase of educational opportunities for minority students in economics	30,753	32,593	(63,346)	-	-
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	-	5,158	(4,090)	-	1,068
The Asia Foundation grant for Asian economists' membership dues to The American Economic Association and related travel expenses	1,442	-	(375)		1,067
The Carnegie Foundation grant for the committee on the status of women in the economics profession	4,866	-	(4,861)		5
The National Science Foundation grant for support of a joint *U.S.-USSR* Symposium on the Economics of Technological Progress	-	11,660	(11,660)	-	-
The Minority scholarship fund for minority students applying for graduate work in economics	-	5,000	-	-	5,000
Sundry	611	100	-	-	711
	$139,676	**$56,878**	**$(146,699)**	**$(1,260)**	**$48,595**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1975

	Balance at January 1	Receipts	Disbursements	Allocation of Investment (Losses) (Note 4)	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics	$256,099	$ 2,074	$(135,661)	$(20,508)	$102,004
The Alfred P. Sloan Foundation, Chase Manhattan Bank, and Ford Foundation grants for increase of educational opportunities for minority students in economics	22,762	40,551	(32,560)	-	30,753
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	6,094	-	(6,094)	-	-
The Asia Foundation grant for Asian economists' membership dues to The American Economic Association and related travel expenses	415	2,871	(1,844)	-	1,442
The Carnegie Foundation grant for the committee on the status of women in the economics profession	20,921	-	(16,055)		4,866
The Kazanjian Foundation grant for the committee on economic education	750	-	(750)	-	-
The German Marshall fund grant for the annual meeting guest speaker		1,500	(1,500)	-	-
Sundry	511	100			611
	$307,552	**$47,096**	**$(194,464)**	**$(20,508)**	**$139,676**

The accompanying notes to financial statements are an integral part of this statement

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN ASSETS AND LIABILITIES FOR THE YEARS ENDED DECEMBER 31, 1976 AND 1975

	1976	1975
Cash (beginning of year)	**$ 121,950**	**$ 37,471**
SOURCE (USE) OF FUNDS:		
Revenues in excess of (less than) expenses	150,375	(55,671)
Add noncash charges		
Depreciation	1,030	983
Directory publication (Note 1)	50,000	50,000
Market value adjustments (Note 1)	37,314	80,678
Funds provided by operations	238,719	75,990
(Increase) decrease in		
Receivables and prepaid expenses	(117,914)	(29,563)
Inventory of *Index of Economic Articles*	(46,098)	-
Investments	(165,116)	97,443
Office furniture and equipment	(644)	(303)
Increase (decrease) in-		
Accounts payable and accrued liabilities	15,879	(100,574)
Deferred income	40,704	72,165
Restricted funds	(91,081)	(167,876)
General fund, market value adjustment	16,288	17,825
Unrecognized change in market value of investments	33,686	119,372
Cash (end of year)	**$ 46,373**	**$ 121,950**

The accompanying notes to financial statements are an integral part of these statements.

NOTES TO FINANCIAL STATEMENTS

(1) Significant Accounting Policies

Investments:

The Association accounts for its investments on a market value basis. Under the method used by the Association to value investments, the change in market value of corporate stocks during the year, after adjusting for an inflation factor (4.7 percent in 1976 and 6.4 percent in 1975), is recognized in income over a three-year period. The change in market value of Treasury Bills, commercial paper, etc., is reflected currently in income. The changes in market value of investments are allocated to the general and restricted fund balances as appropriate.

Accrual for Directory.

Approximately every three to five years, the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was last published in 1974 and distributed at no cost to the membership. In order to match more properly the publishing cost of this directory with revenue from membership dues, the Association provided $50,000 in 1976 and 1975 for estimated publishing costs which will reduce actual directory expense in the year of publication.

Deferred Income:

Revenue from membership dues and subscriptions to the various periodicals of the Association are deferred when received; these amounts are then recognized as income as publications are mailed to the members and subscribers.

Life membership dues are also deferred when received; income is recognized over the estimated average life of these members.

(2) Investments and Investment Income

The following is a summary of investments held by the Association at December 31:

	1976		1975	
	Cost	Market	Cost	Market
Treasury Bills, commercial paper, etc.	$549,081	$549,081	$298,000	$298,000
Corporate stocks	204,943	347,766	312,064	433,731
Total	**$754,024**	**$896,847**	**$610,064**	**$731,731**

Investment losses recognized in income for the years ended December 31, were as follows:

	1976	1975
Treasury Bills, commercial paper, etc.		
Interest	$ 24,780	$ 22,449
Change in market value	-	-
	24,780	**22,449**
Corporate stocks		
Cash dividends	12,284	15,019
Decline in market value recognized (Note 3)	(46,702)	(113,311)
	(34,418)	**(98,292)**
Investment losses allocated to a restricted fund (Note 4)	**5,357**	**27,718**
Investment losses included in income	**$ (4,281)**	**$ (48,125)**

(3) Unrecognized Change in Market Value of Investments

As described more fully in Note 1, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affect 1976 and 1975 revenues, and the amount of these market value increases (declines) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change December 31	
	1976	1975	1977	1978	1976	1975
1973	$ -	$ (58,509)	$ -	$ -	$ -	$ -
1974	(83,715)	(83,715)	-	-	-	(83,715)
1975	28,913	28,913	28,914	-	28,914	57,827
1976	8,100	-	8,099	8,099	16,198	-
	$(46,702)	**$(113,311)**	**$37,013**	**$8,099**	**$45,112**	**$(25,888)**

Included in the above unrecognized changes as of December 31, are increases (declines) of $9,067 and ($7,456) in 1976 and 1975, respectively, which have been allocated to a restricted fund. The amounts allocated are based on the percentage of the Association's total stock portfolio owned by this restricted fund

(4) Restricted Fund

In 1968, the Association entered into an agreement with the University of Colorado relating to the Ford Foundation grant for the Economics Institute which provides, among other things, that the Association invest a portion of the funds received and allocate any income and market value adjustments therefrom to the restricted fund. In accordance with this agreement, the following adjustments were allocated to the restricted fund:

	1976	1975
Net investment losses (Note 2)	$(5,357)	$(27,718)
Market value adjustments arising from inflation	4,097	7,210
Total	**$(1,260)**	**$(20,508)**

(5) Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was $14,625 and $11,910 for 1976 and 1975, respectively.

(6) The Association

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under section 501(c)(3) of the U.S. Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists.

The Association has been determined to be an organization which is not a private foundation.

EXHIBIT I—THE AMERICAN ECONOMIC ASSOCIATION
STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES
FOR THE YEARS ENDED DECEMBER 31, 1976 AND 1975

	1976	1975
Mailing list file maintenance and periodic mailing expenses	$17,344	$14,658
Accounting and legal	10,770	7,600
Office supplies	13,744	10,942
Postage	7,592	11,863
Dues and subscriptions	2,590	2,819
Telephone	3,685	2,802
Investment counsel and custodian fees	2,979	2,780
President and president-elect expenses	3,525	3,000
Travel and entertainment	566	1,732
Depreciation (straight-line method)	1,030	983
Uncollectible receivables	789	51
Currency exchange charges (credits)	(620)	1,443
Insurance and miscellaneous	3,032	1,770
	$67,026	**$62,443**

NOTES

NINETIETH ANNUAL MEETING OF THE AMERICAN ECONOMIC ASSOCIATION

New York, New York, December 27-30, 1977

Preliminary Announcement of the Program

Tuesday, December 27, 1977

10.00 A.M. EXECUTIVE COMMITTEE MEETING

Wednesday, December 28, 1977

8.15 A.M. ECONOMICS OF LIFE AND SAFETY*
Presiding: PHILIP J. COOK, Duke University
Papers: HENRY GRABOWSKI AND JOHN VERNON, Duke University
Consumer Product Safety
THOMAS C. SCHELLING, Harvard University
Willpower and Overeating
AMOS TVERSKY, Hebrew University
Risky Decisions Involving Human Lives
Discussants. ROLAND MCKEAN, University of Virginia
JOHN PRATT, Harvard University
JAMES VAUPEL, Duke University

8.15 A.M. ENERGY AND ECONOMIC GROWTH*
Presiding: LESTER B. LAVE, Carnegie-Mellon University
Papers: EDWARD A. HUDSON, Data Resources Inc., AND DALE JORGENSON, Harvard University
The Relation Between Energy and Economic Growth
WILLIAM D NORDHAUS, Council of Economic Advisers and Yale University
Programming Models for Analyzing the Relation Between Energy and Economic Growth
Discussants TJALLING KOOPMANS, Yale University
CLARK BULLARD, University of Illinois
WILLIAM HOGAN, Stanford University

8.15 A.M. EARNINGS AND EMPLOYMENT OF WOMEN AND RACIAL MINORITIES*
Presiding: FINIS WELCH, University of California-Los Angeles
Papers. MARY CORCORAN, University of Michigan
The Structure of Female Wages
JAMES SMITH, RAND Corporation
Recent Evidence on Black-White Differences in Male Wages
Discussants: GLENN LOURY, Northwestern University
RONALD OAXACA, University of Arizona

8.15 A.M. THE NEGATIVE INCOME TAX AND THE ROLE OF SOCIAL EXPERIMENTATION
Presiding: HAROLD WATTS, Columbia University
Papers: TERRY JOHNSON, PHILIP K. ROBINS, RICHARD W. WEST, Stanford Research Institute, AND JOHN PENCAVEL, Stanford Research Institute and Stanford University
Measuring the Labor Supply Response to a Negative Income Tax Using Results from the Seattle and Denver Income Maintenance Experiments
ROBERT A. MOFFITT, Mathematica Policy Research Inc.
Estimating Labor Supply Response to a Negative Income Tax Experiment in the Presence of Other Tax and Transfer Programs
LYLE P. GROENEVELD, Stanford Research Institute, MICHAEL HANNAN, AND NANCY BRANDON TUMA, Stanford Research Institute and Stanford University
Measuring the Effects of a Negative Income Tax on Marital Status Using Results from the Seattle and Denver Income Maintenance Experiments
Discussants: GUY ORCUTT, Yale University
MILTON FRIEDMAN, University of Chicago and Hoover Institution

8:15 A.M. ANALYTICAL MODELS FOR ASSESSING HOUSING POLICY ECONOMICS
Presiding: FRANK DE LEEUW, U.S. Congressional Budget Office
Papers: JOHN F. KAIN, Harvard University, AND WILLIAM APGAR, National Bureau of Economic Research
Simulation of Housing Market Dynamics
CAROL CORRADO, Federal Reserve Board, AND THOMAS COOLEY, University of California-Santa Barbara
Controlling Fluctuations in Housing Construction: An Analysis of Costs and Policy Options
Discussants: JAMES L. SWEENEY, Stanford University
(Others to be announced)

8:15 A.M. PRICING IN PUBLIC UTILITIES: RECENT ADVANCES IN THEORY AND PRACTICE
Presiding: WILLIAM J. BAUMOL, Princeton University and New York University
Papers: ROBERT D. WILLIG AND ELIZABETH E. BAILEY, Bell Laboratories
Methods for Public Interest Pricing
BRIDGER MITCHELL, RAND Corporation
Pricing in European Public Utilities
A. MICHAEL SPENCE, Harvard University
A Survey of Quantity-Dependent Pricing Models

8:15 A.M. TIME SERIES IN ECONOMIC DEMOGRAPHY
Presiding: T. PAUL SCHULTZ, Yale University
Papers: RICHARD A. EASTERLIN AND MICHAEL WACHTER, University of Pennsylvania
The Passing of the Kuznets Cycle: Is There Life After Death?
ROBERT T. MICHAEL, Stanford University and National Bureau of Economic Research
Causes of the Recent Rise in U.S. Divorce Rates
WILLIAM P. BUTZ, RAND Corporation, AND MICHAEL P. WARD, University of California-Los Angeles
The Emergence of Countercyclical U.S. Fertility
RONALD D. LEE, University of Michigan
A Stock Adjustment Model of U.S. Marital Fertility: 1947–Present
Discussant ROBERT J. WILLIS, Stanford University and National Bureau of Economic Research

8:15 A.M. MONETARY ECONOMICS (Contributed Paper Session)
Presiding: RONALD MCKINNON, Stanford University
Papers: JOHN M. FINKELSTEIN, Indiana University
The Macroeconomic Impact of a Loan Reserve Requirement
DOUGLAS FISHER, North Carolina State University, AND JEFFREY I. BERNSTEIN, Concordia University
Consumption, the Term Structure of Interest Rates, and the Demand for Money
PAUL G. REINHARDT, York University
Inflation-Induced Redistributions: The Wage-Lag Hypothesis Revisited
J. HUSTON MCCULLOCH, National Bureau of Economic Research
Misintermediation and Business Fluctuations
B. R. KOLLURI AND R. SINGAMSETTI, University of Hartford
Permanent Income, Inflationary Expectations and the Demand for Money in U.S.: An Errors-in-Variables Approach

10:30 A.M. CRITIQUE OF OUR SYSTEM*
Presiding ROBERT SOLOW, Massachusetts Institute of Technology
Papers: SAMUEL BOWLES, University of Massachusetts
(Title to be announced)
JAMES BUCHANAN, Virginia Polytechnic Institute and State University
(Title to be announced)
ROBERTO M. UNGER, Harvard University
(Title to be announced)
Discussants: ROBERT SOLOW, Massachusetts Institute of Technology
(Others to be announced)

10:30 A.M. ECONOMICS OF CRIME: DETERRENCE
Presiding: ALFRED BLUMSTEIN, Carnegie-Mellon University
Papers: STEPHEN A. HOENACK, WILLIAM WEILER, AND DAVID L. SJOQUIST, University of Minnesota
Estimates of a Structural Model of Murder Behavior and the Criminal Justice System

CHARLES MANSKI, Carnegie-Mellon University
On the Feasibility of Inferring Deterrent Effects from Observations on Individual Criminal Behavior
ROBERT GORDON, University of Chicago and National Opinion Research Center
An Econometric Model for Measuring the Impact of Legalized Gambling
DANIEL NAGIN, Duke University
Crime Ratios, Sanction, Levels, and Constraints on Prison Populations
Discussants: MICHAEL BLOCH, Stanford University
WALTER VANDAELE, Harvard University

10:30 A.M. INTERNATIONAL TRADE AND THE DEVELOPING COUNTRIES*
Presiding: GOTTFRIED HABERLER, American Enterprise Institute
Papers: CARLOS DIAZ-ALEJANDRO, Yale University
International Markets for LDC's: The Old and the New
RONALD FINDLAY, Columbia University
Direct Foreign Investment and the Transfer of Technology
ANNE O. KRUEGER, University of Minnesota
Alternative Trade Strategies and Employment in LDC's
Discussants: GARY HUFBAUER, U.S. Treasury Department
RONALD MCKINNON, Stanford University
ROBERT BALDWIN, University of Wisconson

10.30 A.M. THE GOALS OF STABILIZATION POLICY*
Presiding. G. LELAND BACH, Stanford University
Papers GARDNER ACKLEY, University of Michigan
The Costs of Inflation
MARTIN S. FELDSTEIN, Harvard University
The Costs of Unemployment
HENRY WALLICH, Board of Governors of the Federal Reserve System
Balancing Inflation and Unemployment
Discussants G LELAND BACH, Stanford University
BARRY BOSWORTH, The Brookings Institution
MICHAEL HURD, Stanford University

10 30 A M. DYNAMIC MODELS OF OLIGOPOLY
Presiding F M. SCHERER, Northwestern University
Papers DARIUS GASKINS, Federal Trade Commission
Dynamic Limit Pricing With Rational Entrants
WAYNE Y. LEE, Indiana University, AND ANDREW J. SENCHACK, JR., University of Texas-Austin
Learning Dynamics and Dominant Firm Behavior
RICHARD SCHMALENSEE, Massachusetts Institute of Technology
Entry Deterrence in the Breakfast Cereal Industry
Discussants CLEMENT G. KROUSE, University of California-Los Angeles
GLENN C. LOURY, Northwestern University

10:30 A.M. INEQUALITY, INCOME DISTRIBUTION, AND INCOME TRANSFERS (Contributed Paper Session)
Presiding MARTIN BRONFENBRENNER, Duke University
Papers. GARY S. FIELDS, Yale University
A Welfare Economic Analysis of Growth and Distribution in the Dual Economy
REBECCA A. MAYNARD, Mathematica Policy Research, Inc., AND RICHARD J. MURNANE, University of Pennsylvania
The Effects of Cash Transfers on School Performance
ROBERT A. MOFFITT, Mathematica Policy Research, Inc.
The Taxation of Earnings in the AFDC Program, 1938–1969
MICHAEL SATTINGER, Miami University
Capital Intensity and Labor Earnings Inequality
JOHN A. TURNER, Social Security Administration
Social Security, Labor Supply and Savings

10 30 A.M. REGULATION IN THE HEALTH CARE INDUSTRY (Joint Session with the Health Economics Research Organization)
Presiding: DONALD E. YETT, University of Southern California

Papers: JERRY CROMWELL, Abt Associates, Inc.
Hospital Rate Regulation
ALAN D. BAUERSCHMIDT, R. W. FURST, AND PHILIP JACOBS, University of South Carolina
An Economic Model of Health Insurance Regulation
JOSEPH LIPSCOMB, Duke University
Prices, Productivity, and the Legal Structure of Dental Practice
Discussants: NANCY O. THORNDIKE, Health Care Financing Administration, U.S. Departmer of Health, Education and Welfare.
LEWIS FREIBERG, JR., National Association of Blue Shield Plans
LEONARD DRABEK, University of Southern California and University of California-Los Angeles

10:30 A.M. ENERGY MODELLING FOR THE U.S. ECONOMY (Joint Session with the Econometric Society)
Presiding: TJALLING KOOPMANS, Yale University
Papers: CLARK BULLARD, University of Illinois
Long Range Energy Demand Analysis
GEORGE DANTZIG, THOMAS CONNOLLY, AND SHAILENDRA PARIKH, Stanford University
PILOT Energy-Economic Model
KENNETH HOFFMAN, DAVID BEHLING, Brookhaven National Laboratory, AND DALE JORGEN SON, Harvard University
Economic and Technological Models for Energy Analysis
ALAN MANNE, Stanford University, AND RICHARD RICHELS, Electric Power Research Institute
A Decision Analysis of the U.S. Breeder Reactor Program
Discussants: AUDIENCE

10:30 A.M. RECENT ADVANCES IN MATHEMATICAL ECONOMICS I (Joint Session with the Econometric Society)
Presiding. KENNETH J. ARROW, Harvard University
Papers: LEONID HURWICZ, University of Minnesota
Resource Allocations Attainable Through Nash Equilibria
HUGO SONNENSCHEIN, Princeton University
Aggregate Demand Functions
JOHN J. MCCALL AND STEVEN A. LIPPMAN, University of California-Los Angeles
Probablistic Economics
Discussants: AUDIENCE

12:30 P.M. JOINT LUNCHEON (With the American Finance Association)
Presiding: LAWRENCE R. KLEIN, University of Pennsylvania
Speaker: W. MICHAEL BLUMENTHAL, Secretary of the Treasury

2:00 P.M. ECONOMICS AND ANTHROPOLOGY: DEVELOPING AND PRIMITIVE ECONOMIES*
Presiding: IRMA ADELMAN, University of Maryland
Panelists: GEORGE DALTON, Northwestern University
CLIFFORD GEERTZ, Princeton University
AMYRA GROSSBARD, Pitzer College

2:00 P.M. LIFE CYCLE AND HOUSEHOLD DECISION MAKING*
Presiding. MARC NERLOVE, Northwestern University
Papers: MARTIN S. FELDSTEIN, Harvard University
The Interdependence of Savings and Retirement Decisions
JAMES J. HECKMAN, University of Chicago
The Dynamics of Female Labor Supply
T. PAUL SCHULTZ, Yale University
Fertility and Mortality in Family Decision Making Over the Life Cycle
Discussants: SHERWIN ROSEN, University of Chicago
ROBERT WILLIS, National Bureau of Economic Research

2:00 P.M. CHANGES IN CONSUMER PREFERENCES*
Presiding: MICHAEL MANOVE, Boston University
Papers: ROBERT A. POLLAK, University of Pennsylvania
Consumer Sovereignty When Tastes are Changing
EDGAR A. PESSEMIER, Purdue University
Stochastic Properties of Changing Preferences
THOMAS A. MARSCHAK, University of California-Berkeley
Economic Issues for Research in Changing Taste
Discussants: RICHARD SCHMALENSEE, Massachusetts Institute of Technology
LESTER D. TAYLOR, University of Arizona

.00 P.M. EFFECTIVENESS OF MONETARY, FISCAL, AND OTHER POLICY TECHNIQUES: COMPETING MEANS*
Presiding: FRANCO MODIGLIANI, Massachusetts Institute of Technology
Papers: ROBERT J. GORDON, Northwestern University
The Effectiveness of Monetary and Fiscal Policies in Controlling Money and Real Incomes
CHARLES C. HOLT, The Urban Institute
Labor Market Policies to Improve the Tradeoff Between Unemployment and Inflation
ARTHUR OKUN, The Brookings Institution
Other Policies to Affect the Unemployment/Inflation Tradeoff
ROBERT LUCAS, University of Chicago
Unemployment Policy

:00 P.M. URBAN AND REGIONAL ECONOMICS (Contributed Paper Session)
Presiding: JULIUS MARGOLIS, University of California-Irvine
Papers: ALAN D. ANDERSON, Carnegie-Mellon University
Long-Run Adjustment of the Urban Housing Stock in Response to Transport Innovation
MICHAEL L. GOETZ AND LARRY E. WOFFORD, University of Tulsa
An Economic Analysis of Land Use Controls
WILLIAM K. HUTCHINSON, Miami University
Regional Exports to Foreign Countries: United States 1870–1910
JOHN D. FILER AND LAWRENCE W. KENNY, University of Florida
Voter Reaction to City-County Consolidation Referenda
PETER LINNEMAN, University of Chicago
The Presence of Children, Location Decisions, and the Suburbanization Process

:00 P.M. RACIAL DISPARITIES AND POLICIES TO ELIMINATE THEM (Joint Session with the National Economic Association)*
Presiding: BERNARD ANDERSON, University of Pennsylvania
Papers: MARCUS ALEXIS, Northwestern University
The Economic Status of Blacks and Whites
HAROLD BLACK AND LEWIS MANDELL, Comptroller of the Currency
Redlining and Discrimination in Mortgage Lending
CHARLES BETSEY, U.S. Department of Labor
Differences in Unemployment Experience Between Blacks and Whites
Discussants: DAVID SWINTON, State University of New York-Stony Brook
RONALD TROSPER, Boston College
KARL GREGORY, Oakland University

:00 P.M. COMPARATIVE RESPONSES TO THE ENERGY CRISIS IN DIFFERENT ECONOMIC SYSTEMS (Joint Session with the Association for Comparative Economic Systems)
Presiding: ARTHUR W. WRIGHT, Purdue University
Papers: Four Case Studies:
YOSHI TSURUMI, Columbia University
—Japan
G. EDWARD SCHUH, Purdue University
—Brazil
JOHN HABERSTROH, Central Intellegence Agency
—Hungary
ARTHUR W. WRIGHT, Purdue University
—United States
Comparative Analyses of the Four Cases:
JUDITH THORNTON AND ROBERT HALVORSEN, University of Washington
JEFFREY B. MILLER, University of Delaware

2:00 P.M. BUREAUCRATIC BEHAVIOR AND CONGRESS (Joint Session with the Public Choice Society)
Presiding: MORRIS FIORINA, California Institute of Technology
Papers: JOHN W. SNOW, U.S. Highway and Traffic Safety Administration
Evaluating the Federal Experience in Motor Vehicle Safety
BARRY WEINGAST, Washington University
The Influence of Congress on Regulatory Decision Making
ANTHONY E. BOARDMAN, University of Pennsylvania
Bureaucratic Behavior and Educational Policy
Discussants: CHARLES GOETZ, University of Virginia
ROBERT GREEN, Carnegie-Mellon University

2:00 P.M. THEORY OF ECONOMIC GROWTH (Joint Session with the Econometric Society)
Presiding: (To be announced)
Papers: RICHARD GILBERT, University of California-Berkeley
Growth and Exhaustible Resources
MUKUL MAJUMDAR, Cornell University
"Turnpike-Type" Results Under Uncertainty
Discussants: (To be announced)

4:00 P.M. ECONOMICS AND LAW*
Presiding: (To be announced)
Papers: WILLIAM LANDES AND RICHARD POSNER, University of Chicago
Economic Analysis of Legal Precedent
KENNETH WOLPIN, Yale University
Upsurge of Crime in Great Britain in the Last 25 Years
CHARLES PHELPS AND RODNEY SMITH, RAND Corporation
The Effects of Regulation on the Oil Industry
Discussant: GUIDO CALABRESI, Harvard University

4:00 P.M. FIRST YEAR OF THE CARTER ADMINISTRATION'S ECONOMIC POLICY (Joint Session with the Society of Government Economists)
Presiding: FRANK SCHIFF, Committee for Economic Development
Papers: (To be announced)

8:00 P.M. RICHARD T. ELY LECTURE*
Presiding: (To be announced)
Speaker: HERBERT A. SIMON, Carnegie-Mellon University
Rationality As Process and as Product of Thought

Thursday, December 29, 1977

8:15 A.M. EFFICIENCY OF MANAGERIAL DECISION PROCESSES*
Presiding: PAUL JOSKOW, Massachusetts Institute of Technology
Papers: JOSEPH BOWER, Harvard University
The Business of Business is Serving Markets
HARVEY LEIBENSTEIN, Harvard University
An X-Inefficiency, Nontradeoff Theory of Firm Organization
SIDNEY WINTER, University of Michigan
On the Economics of Attention
Discussants: RICHARD DAY, University of Southern California
MICHAEL ROTHSCHILD, University of Wisconsin
(Others to be announced)

8:15 A.M. ENERGY POLICY
Presiding: WALTER J. MEAD, University of California-Santa Barbara
Papers: WALTER J. MEAD, University of California-Santa Barbara
An Overview of Past U.S. Energy Policies
MILTON RUSSELL, Resources for the Future
U.S. Energy Policy Options
STEPHEN L. MCDONALD, University of Texas-Austin
U.S. Energy Resource Leasing Policy
HENDRIK S. HOUTHAKKER, Harvard University
U.S. International Energy Policy
Discussants: PAUL W. MACAVOY, Yale University
JOHN W. WILSON, Consulting Economist

8:15 A.M. ECONOMIC EDUCATION*
Presiding: G. LELAND BACH, Stanford University
Papers: DAVID HARTMAN, Harvard University
What Do Majors in Economics Learn?

Discussants: ROUND TABLE: "WHAT DO (SHOULD) MAJORS IN ECONOMICS LEARN?"
MARTIN BRONFENBRENNER, Duke University
BARBARA REAGAN, Southern Methodist University
ROBERT SOLOW, Massachusetts Institute of Technology
ROBERT THOMAS, University of Washington

8:15 A.M. DEVELOPMENT, PLANNING, AND GROWTH (Contributed Paper Session)
Presiding: HUEY J. BATTLE, Virginia State College
Papers. ANTONIO MARIA COSTA, United Nations
International Linkage of Econometric Models for the CMEA Region: Characteristics and Policy Responses
JAMES A. EDMONDS, Centre College of Kentucky and Institute for Energy Analysis
The Interaction of Economic Growth, Population Growth, and Human Capital
CHARLES MICHAEL ELLIS, University of Central Arkansas
Regional Economic Development and Income Inequality: A Principal Component-Canonical Correlation Analysis
DONALD A. HANSON, Southern Methodist University
Efficient Transitions From a Resource to a Substitute Technology in a Macroeconomic Growth Context
GIAN S SAHOTA, Vanderbilt University, AND CARLOS A. ROCCA, University of Sao Paulo
Distribution in the Process of Growth

8 15 A M. HUMAN RESOURCES AND LABOR SUPPLY (Contributed Paper Session)
Presiding T. PAUL SCHULTZ, Yale University
Papers BARRY T HIRSCH, University of North Carolina-Greensboro
Earnings Inequality Across Labor Markets: A Test of the Simple Human Capital Model
MARK R. ROSENZWEIG, Princeton University
Neoclassical Theory and the Optimizing Peasant. An Econometric Analysis of Labor Supply in a Developing Country
STEVEN H. SANDELL, Ohio State University
Is the Unemployment Rate of Women Too Low? A Direct Test of the Economic Theory of Job Search
SHARON P. SMITH AND JAMES F. RAGAN, Federal Reserve Bank of New York
The Impact of Differences in Turnover Rates on Male/Female Pay Differentials
STANLEY P STEPHENSON, JR, Pennsylvania State University
The School to Work Transition of Young Men Aged 16-24

8 15 A M. REGULATION, LAW, AND EXTERNALITIES (Contributed Paper Session)
Presiding WILLIAM M. LANDES, University of Chicago
Papers: MELVIN J. HINICH, Virginia Polytechnic Institute and State University, AND RICHARD STAELIN, Carnegie-Mellon University
A Study of the Effects of Food Regulation
KEVIN HOLLENBECK, Mathematica Policy Research, Inc.
The Employment and Earnings Incidence of the Regulation of Air Pollution
SAMUEL L. MYERS, University of Texas-Austin
An Economic Model of Black-White Differentials in the Post-Release Behavior of Criminal Offenders
GEORGE H. SWEENEY, Vanderbilt University
Dynamic Behavior of a Firm Subject to Rate of Return Regulation
DONALD WITTMAN, University of California-Santa Cruz
Optimal Methods of Monitoring and Control with Special Reference to the Legal System

8 15 A M INDUSTRIAL LOCATION DECISIONS (Joint Session with the Econometric Society)
Presiding: WALTER ISARD, University of Pennsylvania
Papers: DONALD ERLENKOTTER, University of California-Los Angeles
Dynamic Models of Industrial Location Decisions
ROGER SCHMENNER, Harvard University, AND DAN WEINBERG, Abt Associates, Inc.
Evolutionary Model of Intra-Metropolitan Location Decisions
DENNIS W. CARLTON, University of Chicago
Econometric Models of New Business Location
Discussants: PETER KEMPER, University of Wisconsin
RAYMOND J. STRUYK, Department of Housing and Urban Development
(Others to be announced)

8:15 A.M. RECENT ADVANCES IN MATHEMATICAL ECONOMICS II (Joint Session with the Econometric Society)
Presiding: MICHAEL D. INTRILIGATOR, University of California-Los Angeles
Papers: LIONEL MCKENZIE, University of Rochester
Optimal Economic Growth and Turnpike Theorems
W. ERWIN DIEWERT, University of British Columbia
Duality Approaches to Microeconomic Theory
ROBERT C. MERTON, Massachusetts Institute of Technology
Investment Theory
Discussants: AUDIENCE

10:30 A.M. ECONOMICS AND ETHICS: ALTRIUSM, JUSTICE, POWER*
Presiding: MORDECAI KURZ, Stanford University
Papers: JOHN HARSANYI, University of California-Berkeley
Utilitarianism, Preferences, Rules, and Rationality
THOMAS C. SCHELLING, Harvard University
Strategically Self-Serving Behavior
MORDECAI KURZ, Stanford University
Altruism as an Outcome of Social Interaction
Discussants: ROBERT TRIVERS, Harvard University
ROGER MYERSON, Northwestern University
JOHN HARSANYI, University of California-Berkeley

10:30 A.M. QUALITY OF WORKING LIFE*
Presiding: LLOYD ULMAN, University of California-Berkeley
Papers. KARL-OLOF FAXÊN, Swedish Employers Confederation
Disembodied Technical Progress: Does Employee Participation in Decision Making Contribute to Change and Growth?
R. B. FREEMAN, Harvard University
Job Satisfaction as an Economic Variable
LESTER C. THUROW, Massachusetts Institute of Technology
The Concept of Psychic Income: Useful or Useless?
Discussants: R. A. OSWALD, Service Employees Int. Union, Washington
MICHAEL J. PIORI, Massachusetts Institute of Technology
GEORGE STRAUSS, University of California-Berkeley

10:30 A.M. ECONOMICS OF DEVELOPMENT. SELF-CRITIQUE FROM TWO SIDES
Presiding: (To be Announced)
Papers: JOHN G. GURLEY, Stanford University
What Has Marxist Economics Learned From Recent Experience?
JEFFREY B NUGENT, University of Southern California, AND PAN A YOTOPOULOS, Stanford University
Has Orthodox Economics Learned from Marxist Approaches?
DAVID MORAWETZ, Hebrew University
The Economics of Development of the Socialist Countries: What Can We Learn From Their Experience?
ALBERT FISHLOW, University of California-Berkeley
The Economics of Development of Some Socialist Developing Countries: What Can Be Learned from their Experience?
Discussants: (To be announced)

10:30 A.M. PROBLEMS OF REGIONAL ECONOMIC DEVELOPMENT*
Presiding: LEON MOSES, Northwestern University
Papers: DAVID KRESGE, National Bureau of Economic Research, AND DANIEL SEIVER, University of Alaska
Planning for a Resource-Rich Region: The Case of Alaska
F. LEE BROWN AND ALLEN V. KNEESE, University of New Mexico
The Southwest: A Region Under Stress
ROBERT A. LEONE AND JOHN R. MEYER, Harvard University
The Northeastern States and Their Future
Discussants: BENJAMIN CHINITZ, State University of New York-Binghamton
WALTER ISARD, University of Pennsylvania

10:30 A.M. COST-BENEFIT ANALYSIS (Contributed Paper Session)
Presiding: ROLAND N. MCKEAN, University of Virginia
Papers: RITA R. CAMPBELL, Hoover Institution
Prospective Cost-Benefit Analysis of a New Drug "Tagamet" (Cimetidine) Used to Treat Duodenal Ulcer
W. MARK CRAIN, Virginia Polytechnic Institute and State University
The Economics of Mandatory State Vehicle Inspections
CHARLES M. GRAY, Governor's Commission on Crime Prevention and Control, St. Paul, MN
Neighborhood Crime and the Demand for Central City Housing
JOHN KRAFT AND MARK RODEKOHR, Federal Energy Administration
An Analysis of the Distribution of Benefits Associated with a Two-Tier Natural Gas Market

10:30 A.M. DECISION MAKING IN THE PRIVATE AND PUBLIC SECTORS (Contributed Paper Session)
Presiding LELAND W. JOHNSON, RAND Corporation
Papers LEE S. FRIEDMAN, University of California-Berkeley
A Public Sector Application of an Evolutionary Model: The Production of Bail Reform
FRANK LEVY AND MARK KAMLET, University of California-Berkeley
The Revealed Preference of a Satisficing Bureaucrat
ROBERT J. MICHAELS, California State University-Fullerton
Public Production in a Declining Industry: The Case of State Mental Hospitals
MICHELLE J. WHITE, University of Pennsylvania
The Economics of Partnership

10:30 A.M. DEMAND AND PRICING (Contributed Paper Session)
Presiding ELIZABETH E. BAILEY, Bell Laboratories
Papers SCOTT E. ATKINSON, Federal Energy Administration
A Preliminary Analysis of the FEA Time-of-Day Electricity Pricing Experiments
ROBERT E. DANSBY, Bell Laboratories
An Economic Evaluation of Interruptible Service Offerings
DAVID S SIBLEY, M. BARRY GOLDMAN, Bell Laboratories, AND HAYNE E. LELAND, University of California-Berkeley
Optimal Non-Uniform Pricing
ALFRED FRANCFORT, University of Pittsburgh-Johnstown, AND J D. STENGER, United Illuminating Company, New Haven, CT
Response of Residential Customers to Time-of-Use Rates for Electricity A Case Study

10:30 A.M. ECONOMICS OF CRIME AND LAW ENFORCEMENT (Joint Session with Omicron Delta Epsilon)
Presiding: MICHAEL SZENBERG, Long Island University
Papers DAVID A. KENNETT, Columbia University
The Distribution of Unpriced Public Goods A Theoretical Overview with Some Empirical Results for Police Services
MARIO RIZZO, University of Chicago
Rent Property Values and Cost of Crime to Victims
EMANUEL HAAS, City University of New York
The Determinants of Municipal Crime Rates and Law Enforcement Expenditures
Discussants: ROGER MEINERS, University of Miami
ALAN COHEN, University of Wisconsin
STU GUTTERMAN, State University of New York-Stony Brook

10:30 A.M. ECONOMIC ANALYSIS OF REGULATION (Joint Session with the Transportation and Public Utilities Group of the AEA)
Presiding WILLIAM G. SHEPHERD, University of Michigan
Papers JAMES R. NELSON, Amherst College
Marginal-Cost Pricing and Related Devices
JOHN B. SHEAHAN, Williams College
Public Enterprise. Ideas From European Experience
PAUL MACAVOY, Yale University
Deregulation: Politics and Economics as Regulatory Reform
Discussant. DONALD J DEWEY, Columbia University

12 30 P M. LUNCHEON HONORING THE 1976 NOBEL LAUREATE MILTON FRIEDMAN
Presiding: SIMON KUZNETS, Harvard University
Speaker: KARL BRUNNER, University of Rochester

2:00 P.M. DECENTRALIZATION, BUREAUCRACY, AND GOVERNMENT*

Presiding: MANCUR OLSON, University of Maryland

Papers. WILLIAM BROCK, University of Chicago AND STEPHEN MAGEE, University of Texas

Economics of Special-Interest Politics

JOEL GUTTMAN, University of California-Los Angeles

"Semi-Cooperative" Equilibria in the Provision of Public Goods

ROGER NOLL AND MORRIS FIORINA, California Institute of Technology

The Electoral Foundations of the Growth of Bureaucracy

Discussants DANIEL MCFADDEN, University of California-Berkeley

RICHARD NELSON, Yale University

WALLACE OATES, Princeton University

2:00 P.M. ECONOMICS AND MORAL VALUES

Presiding: KENNETH E. BOULDING, University of Colorado

Panelists. CAROLYN SHAW BELL, Wellesley College

EMILE BENOIT, Columbia University

MATTHEW EDEL, Queens College, City University of New York

NEIL H. JACOBY, University of California-Los Angeles

JOHN C O'BRIEN, California State University-Fresno

2:00 P.M. LOCAL REGULATION AND THE ECONOMY OF THE CITY

Presiding HAROLD HOCHMAN, City University of New York

Papers: MICHELLE WHITE, University of Pennsylvania

No Growth in the Suburbs. A New Zoning Strategy

ELIZABETH ROISTACHER, Queens College, AND JAMES SUAREZ, Hunter College

Local Regulation in New York City. Rent Controls and Qualitative Controls

ROBERT INMAN, University of Pennsylvania, AND DANIEL RUBINFELD, University of Michigan

Legal Regulations, Equity, and the Local Public Sector

Discussants BENNETT HARRISON, Massachusetts Institute of Technology

DON MARTIN, University of Miami

2:00 P.M. INTERNATIONAL ECONOMICS (Contributed Paper Session)

Presiding: JOHN PIPPENGER, University of California-Santa Barbara

Papers JAY H. LEVIN, Wayne State University

Devaluation, the J-Curve, and Flexible Exchange Rates

RACHEL MCCULLOCH, Harvard University, AND JANET L. YELLEN, Board of Governors of the Federal Reserve System

Technology Transfer and the National Interest

ANDREW STERN, California State University-Long Beach

Alternative Liability Assignments and Economic Integration

EDWARD JOHN RAY, Ohio State University

Optimal Intervention in a Dynamic Setting

2:00 P.M. MACROECONOMICS (Contributed Paper Session)

Presiding: DONALD BEAR, University of California-San Diego

Papers. JOHN DAVID FERGUSON AND WILLIAM R. HART, Miami University

Liquidity Preference or Loanable Funds: An Empirical Analysis of Interest Rate Determination in Market Disequilibrium

HERBERT M. KAUFMAN, Arizona State University

An Extension of the "Crowding Out" Hypothesis to Federal Agency Activity

R. F. LUCAS, University of Saskatchewan

Tariffs, Nontraded Goods and the Optimal Stabilization Policy

MACK OTT AND ROBERT M. SPANN, Virginia Polytechnic Institute and State University

Variations on the Fisherian Theme: Uncertainty, Inflation, Corporate Financing Decisions, and the Term Structure of Interest Rates

2:00 P.M. EVALUATION OF ENVIRONMENTAL DAMAGE (Joint Session with the Association of Environmental and Resource Economists)

Presiding: ALLEN V. KNEESE, University of New Mexico

Papers: RALPH C. D'ARGE, University of Wyoming, AND WILLIAM D. SCHULZE, University of Southern California

Valuing Recreational Damages

EDWIN S. MILLS, Princeton University
Valuing Water Quality
V. KERRY SMITH, Resources for the Future
Valuing Preservation of Unique Areas in View of Recent Natural Resources Developments
Discussants: CHARLES J. CICCHETTI, University of Wisconsin
ALAN RANDALL, University of Kentucky
SUSAN ROSE-ACKERMAN, Yale University
BURTON A. WEISBROD, University of Wisconsin
RONALD G. CUMMINGS, University of New Mexico
JOHN V. KRUTILLA, Resources for the Future

2:00 P.M. LABORATORY EXPERIMENTAL METHODS IN POLITICAL ECONOMY (Joint Session with the Public Choice Society)
Presiding: RICHARD MCKELVEY, Carnegie-Mellon University
Papers: CHARLES R. PLOTT AND J. HONG, California Institute of Technology
Implications of Rate Filing for Transportation on Inland Waters: An Experimental Approach
JOHN KAGEL, Texas A&M University
The Laws of Demand With Animal Consumers
VERNON L. SMITH, University of Arizona
An Unanimity Process for Private, Public and Externality Goods
Discussants: RICHARD E KIHLSTROM, University of Illinois
STANLEY REITER, Northwestern University

2:00 P.M. TRANSPORTATION RESEARCH (Joint Session with the Transportation and Public Utilities Group of the AEA)
Presiding: ROBERT D. PASHEK, Pennsylvania State University
Papers: ERNEST R. OLSON, Interstate Commerce Commission
Research in Transportation Economics—The Unfulfilled Promise
JOSEPH L. CARROLL AND KANT RAO, Pennsylvania State University
Economics of Public Investment in Inland Navigation: Unanswered Questions
Discussants: (To be announced)

4:00 P.M. HOW HAVE FORECASTS WORKED?*
Presiding: LAWRENCE R. KLEIN, University of Pennsylvania
Papers: PHOEBUS DHRYMES, Columbia University
Forecasting Performance of Selected Econometric Models of the United States, 1971-74
STEPHEN MCNEES, Federal Reserve Bank of Boston
The "Rationality" of Economic Forecasts
VINCENT SU, City University of New York
An Error Analysis of Econometric and Non-Econometric Forecasts
VICTOR ZARNOWITZ, University of Chicago
On the Accuracy and Properties of Recent Macroeconomic Forecasts
Discussants: OTTO ECKSTEIN, Harvard University
ARTHUR OKUN, The Brookings Institution

8:00 P.M. PRESIDENTIAL ADDRESS
Presiding: (To be announced)
Speaker: LAWRENCE R. KLEIN

9:00 P.M. BUSINESS MEETING

Friday, December 30, 1977

8:15 A.M. PSYCHOLOGY AND ECONOMICS*
Presiding: DAVID M. GRETHER, California Institute of Technology
Papers. JAMES MORGAN, University of Michigan
Economic Decisions: The Psychology and Economics of Uncertainty, Ignorance, and Confusion
PAUL SLOVIC, Decision Research, Eugene, Oregon, AND HOWARD KUNREUTHER, University of Pennsylvania
Human Behavior and Public Policy: Implications of Laboratory and Field Research
DAVID M. GRETHER, California Institute of Technology
On Some Recent Psychological Studies of Behavior Under Uncertainty

Discussants: George Katona, University of Michigan
Vernon Smith, University of Arizona
Amos Tversky, Hebrew University

8:15 A.M. Economics of Education
Presiding: W. Lee Hansen, University of Wisconsin
Papers: James W. Albrecht, Columbia University
Interpreting the Returns to Education
John H. Bishop, University of Wisconsin
Impact of Public Policy on College Attendance
W. Lee Hansen, H. B. Neuberger, F. J. Schroeder, B. A. Weisbrod, University of Wisconsin-Madison, D. C. Stapleton, University of British Columbia, and D. J. YoungDay, Montana State University
Graduate Education and the Market for Ph.D. Economists
Discussants: Stephen Dresch, Institute for Demographic and Economic Studies
Greg J. Duncan, University of Michigan
Charlotte V. Kuh, Harvard University

8:15 A.M. Housing Allowances Experiments: Early Findings
Presiding: Donna Shalala, Department of Housing and Urban Development
Papers Ira S. Lowry, RAND Corporation
Early Findings of the Supply Experiment
Steven Kennedy, Abt Associates, Inc
Economics of Housing Demand in the Demand Experiment
John D. Heinberg, The Urban Institute
Integrated Analysis of Housing Allowances: Synthesizing a Complex Research Program
Discussants. John F. Kain, Harvard University
Katherine Lyall, Department of Housing and Urban Development
Guy Orcutt, Yale University

8:15 A.M. The Development of Economics. The Last Fifty Years (Joint Session with the History of Economics Society)
Presiding Warren J. Samuels, Michigan State University
Papers Sir Eric Roll, Warburg & Co., London, England
Economics in Historical Perspective
Robert Heilbroner, New School for Social Research
The Development of Economics. The Last Fifty Years
Discussants: John Kenneth Galbraith, Harvard University
Craufurd Goodwin, Duke University
William O. Thweatt, Vanderbilt University

8:15 A.M. Unemployment in Comparative Perspective (Joint Session with the Association for Comparative Economic Studies)*
Presiding: Franklin D. Holzman, Tufts University
Papers: Morris Bornstein, University of Michigan
Unemployment in Capitalist Market Economies and Socialist Planned Economies
Robert Haveman, University of Wisconsin
Structural Factors, Aggregate Demand, and Unemployed Resources: Western Europe and the United States
Henry J. Bruton, Williams College
Unemployment Problems and Policies in Less Developed Countries
Discussants: Janet G. Chapman, University of Pittsburgh
Nancy Smith Barrett, American University
Howard Pack, Swarthmore College

8:15 A.M. Mathematical Models in Political Science and Economics (Joint Session with the Public Choice Society)
Presiding: Robert Parks, Washington University
Papers: Linda Cohen, California Institute of Technology
The Structure of Maximum Majority Rule Cycles
Ken Shepsle, Washington University
Theories of Legislative Rules

NORMAN SCHOFIELD, University of Texas-Austin
Local Acyclicity Under Arbitrary Social Preference Functions
RICHARD MCKELVEY, Carnegie-Mellon University
Methods for Determining Optimum Majority Rule Agendas

8:15 A.M. ECONOMIC TIME SERIES (Joint Session with the Econometric Society)
Presiding: (To be announced)
Papers: BENOIT MANDELBROT, IBM Research Center
Cyclic Non-Periodic Aspects of Economic Time Series
PETER ROBINSON, Harvard University
Distributed Lag Approximations to Linear Time-Invariance Systems
Discussants: (To be announced)

8:15 A.M. MARKETS UNDER UNCERTAINTY (Joint Session with the Econometric Society)
Presiding: (To be announced)
Papers: JERRY GREEN, Harvard University
On the Value of Information in Incomplete Market Systems
JIM JORDAN, University of Minnesota
Recent Developments in Expectations Equilibrium Theory
ROY RADNER, University of California-Berkeley
Rational Expectations with Price Information
Discussants: (To be announced)

8:15 A.M. DEMOGRAPHIC DETERMINANTS OF PRIVATE SAVING
Presiding: GEORGE M. VON FURSTENBERG, Indiana University
Papers. PAUL WACHTEL, New York University
Will the Changing Age Structure Depress Saving?
THOMAS E. MACURDY, National Bureau of Economic Research, AND ROBERT SCHMITZ, University of Chicago
Lifetime Labor Supply and Interdependence Between Human and Nonhuman Asset Accumulation
SUSAN W. BURCH, Board of Governors of the Federal Reserve System
Microeconomic Saving Determinants Revealed From Surveys of Consumer Expenditures and Financial Assets
Discussants: ALAN S. BLINDER, Princeton University
JOHN W. GRAHAM, University of Illinois-Urbana

10.30 A.M. ECONOMICS AND BIOLOGY: EVOLUTION, SELECTION, AND THE ECONOMIC PRINCIPLE*
Presiding: RICHARD R. NELSON, Yale University
Papers. MICHAEL T. GHISELIN, University of California-Berkeley
The Economy of the Body
ROBERT TRIVERS, Harvard University
The Bio-Economics of the Family
JACK HIRSHLEIFER, University of California-Los Angeles
Cooperation, Conflict, and Competition in Economics and Biology
Discussants GARY S. BECKER, University of Chicago
RONALD H. COASE, University of Chicago
ERIC CHARNOV, University of Utah

10 30 A.M. THE EFFECTS OF THE INCREASED LABOR FORCE PARTICIPATION OF WOMEN ON MACROECONOMIC GOALS*
Presiding. BARBARA REAGAN, Southern Methodist University
Papers: BETH NIEMI, Rutgers University, AND CYNTHIA B. LLOYD, Barnard College
The Effects of Economic and Demographic Change on Sex Differentials in Labor Supply Elasticity
BARBARA BERGMANN AND KATHERINE SWARTZ, University of Maryland, AND CLAIR VICKERY, University of California-Berkeley
Old Policies, New Policies, or No Policies for Dealing With Women's High Unemployment Rates
R. CHRISTOPHER LINGLE, West Georgia College, AND ETHEL B. JONES, Auburn University
Women's Increasing Unemployment: A Cross-Sectional Analysis
Discussants: BARRY R. CHISWICK, Hoover Institution
RALPH E. SMITH, The Urban Institute
MARIANNE A. FERBER, University of Illinois

10:30 A.M. CONGRESSIONAL SUPERVISION OF MONETARY POLICY
Presiding: KARL BRUNNER, University of Rochester
Panelists: ALLAN H. MELTZER, Carnegie-Mellon University
JAMES PIERCE, University of California-Berkeley
ROBERT WEINTRAUB, House Banking and Currency Committee
PAUL VOLCKER, Federal Reserve Bank of New York
Discussants: AUDIENCE

10:30 A.M. THE ECONOMICS OF INFORMATION (Contributed Paper Session)
Presiding: FRITZ MACHLUP, Princeton University
Papers: W. MICHAEL COX, Virginia Polytechnic Institute and State University
Rational Expectations and the Monetary Approach to the Balance of Payments
PETER HOWITT, University of Western Ontario
Activist Monetary Policy Under Rational Expectations
MACK OTT, Virginia Polytechnic Institute and State University, AND STEVEN MILLSAPS, Appalachian State University
OCS Oil Leasing, Consortia Formation and Bidding Behavior: An Application of the Theory of Risk Aversion
F. OWEN IRVINE, JR., Wesleyan University
An Optimal Firm Price Adjustment Policy. The "Short-Run Inventory Based Pricing Policy" Hypothesis

10:30 A.M. U.S. ECONOMIC POLICY IN BLACK AFRICA AND THE CARIBBEAN (Joint Session with the National Economics Association)
Presiding: VINCENT MCDONALD, Howard University
Papers: DIANNE PAINTER, Wellesley College
U.S. Economic Policy in Black Africa
RANSFORD PALMER, Howard University
The U.S. Economy and Caribbean Dependence
ALEX WILLIAMS, University of Virginia
External Debt and Economic Growth in Developing Countries and Implications for U.S. Economic Policy, with Particular Emphasis on Black Africa
Discussants: RAWLE FARLEY, State University of New York-Brockport
LASCELLES ANDERSON, University of Akron
GLENN LOURY, Northwestern University

10:30 A.M. INTERNATIONAL EXCHANGE RATES AND THE MACROECONOMICS OF OPEN ECONOMIES (Joint Session with the American Finance Association)*
Presiding: MARINA V. N. WHITMAN, University of Pittsburgh
Papers: JOHN BILSON, Northwestern University and International Monetary Fund
The Current Experience With Floating: A Monetary View
PETER KENEN, Princeton University
New Views of Exchange Rates and Old Views of Policies
NORMAN MILLER, University of Pittsburgh
International Money Flows and National Money Supplies
Discussants: RUDIGER DORNBUSCH, Massachusetts Institute of Technology
JACOB FRENKEL, University of Chicago
MARC MILES, Rutgers University

10:30 A.M. SLOWDOWN IN THE GROWTH OF PRODUCTIVITY IN THE UNITED STATES (Joint Session with the American Finance Association)
Presiding. M. ISHAQ NADIRI, National Bureau of Economic Research
Papers: JOHN W. KENDRICK, U.S. Department of Commerce
Total Investment and Productivity Developments
PETER K. CLARK, Council of Economic Advisers
Capital Formation and the Recent Productivity Slowdown
MICHAEL D. MCCARTHY, Wharton Econometric Forecasting
Current Prospects for Productivity Growth
GEORGE M. VON FURSTENBERG, Indiana University
The Relation Between Government Debt and Capital Intensiveness in the Long Run
Discussants. EDWARD F. DENISON, The Brookings Institution
J. RANDOLPH NORSWORTHY, U.S. Department of Labor

10:30 A.M. DECENTRALIZATION, TAXATION, AND THE REVELATION OF PREFERENCES FOR PUBLIC DECISIONS (Joint Session with the Econometric Society)
Presiding: ERWIN DIEWERT, University of British Columbia
Papers: JEAN-JACQUES LAFFONT, Harvard University
Revelation of Preferences and Private Information
EYTAN SHESHINSKI, Harvard University
Optimal Taxation and Public Goods
Discussants: KENNETH ARROW, Harvard University
PETER DIAMOND, Massachusetts Institute of Technology

10:30 A.M. CONTEMPORARY ISSUES IN NATURAL RESOURCE ECONOMICS (Joint Session with the American Agricultural Economics Association)
Presiding. EMERY N. CASTLE, Resources for the Future, Inc.
Papers: JOHN FEREJOHN AND TALBOT PAGE, California Institute of Technology
Intertemporal Social Choice: Long Range Decision Making
RICHARD DAY, University of Southern California
Adaptive Economics and Natural Resource Policy
V. KERRY SMITH, Resources for the Future, Inc.
Scarcity and Growth Revisited
Discussants: GEORGE TOLLEY, University of Chicago
ALLEN RANDALL, University of Kentucky

Editor's Note:
*Papers from sessions marked with an asterisk will be published in the *Papers and Proceedings* issue of the *Review*.

ANNOUNCEMENTS

The ninetieth annual meeting of the American Economic Association will be held in New York City, December 28-30, 1977. The Employment Center will be open from December 27-30.

Annual Meeting Employment Center

The Employment Center at the 1977 annual meetings of the Allied Social Science Associations in New York City will begin operation on December 27, the day before sessions begin. Applicants and employers will be able to attend more sessions with a day set aside entirely for labor market transactions. This service will be located at the Convention Employment Center in the Americana Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28-29; and 9:00 A.M. to 12:00 noon, December 30.

Because the 1978 annual meeting comes at an early date in the academic year (August 29-31), the Executive Committee has decided to provide employment services at a later time. A job market will *not* be organized for the August meeting in Chicago. The AEA will provide an organized market in December 1978 or January 1979 at a site yet to be selected. Only one placement service will be provided during the academic year 1978-79, and it will be separate from and later than the annual meeting

---

Call for Papers for the 1978 Meetings

Members wishing to give papers or make suggestions for the program for the meetings to be held in Chicago, August 29-31, 1978, are invited to send their ideas to Professor Robert Solow, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. Although most of the sessions sponsored by the American Economic Association will consist of invited papers, there will also be several sessions of noneconometric contributed papers. (The sessions of contributed papers will not be published in the *Papers and Proceedings* issue to appear February 1979.) Proposals for invited sessions should be submitted as soon as possible. To be considered for the contributed sessions, abstracts of proposed (non-econometric) papers must be received no later than February 1, 1978. Economists wishing to give papers on econometrics or economic theory may submit abstracts to the Econometric Society, which meets with the American Economic Association and annually schedules a substantial number of contributions.

The Law and Economics Center of the University of Miami School of Law is accepting applications to the John M. Olin Fellowship Program in Law and Economics for the class entering in September 1978. The three-year program, designed for highly motivated individuals with a strong foundation in graduate microeconomics and its application, leads to the Juris Doctor degree and prepares Fellows for scholarly research, teaching, legal practice, or government work in the various fields of Law and Economics. Fellows who have passed the preliminary examinations for the Ph.D. in economics are encouraged to complete the dissertation during their residence in Miami. Fellows have the option to complete the program in two and a half years by attending summer school.

Five fellowships of approximately $30,000 each (roughly $10,000 per year including tuition and fees) are available each year. Deadline for submitting applications is February 15, 1978. Awards will be announced by April 1, 1978. Candidates must apply separately to the School of Law, and should plan to take the Law School Admissions Test in December 1977 or no later than February 1978. Individuals interested in more detailed information should write to the Coordinators, John M. Olin Fellowship Program, Law and Economics Center, University of Miami School of Law, P.O. Box 248000, Coral Gables, Florida 33124.

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies Grants are likely to cover only lowest cost excursion fares and will rarely exceed 50 percent of full economy-class fares. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are ineligible. The deadlines for applications to be received in the office of the American Economic Association are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212.

The U.S. Information Agency (*USIA*) has in recent years encouraged economists to meet with, or speak to, groups of economists when traveling abroad. The general subject areas in which American "volunteer speakers" have been encouraged to give talks or participate in symposia in all regions of the world include: international cooperation in management of the inter-

national economy; international economic cooperation—relations between the industrialized and developing countries; American direct investment abroad and the multinational corporation; international trade; international finance and the international monetary system; and topics or subtopics with a somewhat lower priority, East-West economic relations; trade preferences for less developed countries; and commodities.

Topics centering on the United States which are given a priority are the American economic system and U.S energy policies—domestic and international. The *USIA* is particularly interested in learning when American economists are traveling abroad who are knowledgeable in these areas of interest, and are also involved in research on the economies of other countries or regions, perhaps including those which are to be visited. Of special interest to *USIA*, also, is whether economists with such expertise are able to lecture and converse in one or more of the languages spoken in countries to be visited.

USIA is also desirous of expanding its programming with economists whose research and writings span the lines between economics and the behavioral sciences. Economists with such area, research, and/or language competence, who are interested in participating in *USIA*'s economic programs abroad or learning more about *USIA*'s programs relating to economic subjects, are invited to write ICS/DE, U S. Information Agency, Room 615, 1717 H St., N.W., Washington, D.C. 20547.

The P W S Andrews Memorial Prize is awarded annually for an essay by a young scholar (under the age of 30 or within 8 years of taking his first degree) in the general field of Industrial Economics and the Theory of the Firm, broadly interpreted The prize is £150 (or the equivalent in other currency) and the winning essay will normally be published in *The Journal of Industrial Economics* An essay submitted should be a work of original research by the candidate only, not previously published and not previously awarded any other prize. It should be submitted in English and should not normally exceed 10,000 words in length The closing date for entries is 31 December in each year Intending candidates for the prize should obtain details of the conditions of entry from the Administrative Officer, Office for Student and College Affairs, University House, University of Lancaster, Lancaster, LA1 4YW, England.

Invest-in-America announces that it is receiving requests for grants to support Institutes on the American Economy and on the Economics of Consumerism for the summer of 1978. These institutes run for three to six weeks depending on the number of credits that are to be granted by the host college or university. They are intended for teachers in the precollege school system. Invest-in-America will make a cash grant to the host school as well as provide it with several speakers at no charge. Requirements for the awarding of a grant include: ability of the host school to attract about forty teachers to the program, cooperation of the schools or departments of education and economics, and handling administrative details. Those interested in applying for such grants should communicate with Dr. Reuben E. Slesinger, Professor of Economics, 313 Mervis Hall, University of Pittsburgh, Pittsburgh, PA 15260. Grants will be announced around January 1978.

National Humanities Center Fellowships

The National Humanities Center will accept twenty-five Fellows for the academic year 1978-79. Fellowships will be available for 1) Senior Scholars who have attained distinction in their fields; 2) Young Scholars of promise who have received their Ph.D 's in the last six to ten years, and 3) Scholars of all ages whose proposed projects are connected with the topics, Man and Nature, History and the History of Ideas, or The Foundations of the American Polity The Center will provide full support to some Fellows at the level of their normal academic salaries and encourages application for residence from Fellows who have full or partial support from other sources. The National Humanities Center admits Fellows of any race, color, or national or ethnic origin. The deadline for applications is December 1, 1977. Further information and applications may be obtained from The National Humanities Center, P O Box 12256, Research Triangle Park, NC 27709

The Division of Social Sciences of the National Science Foundation announces a Law and Social Sciences Program to support basic social science research on the operation, impact, and use and change of legal and law-like systems of social control. The current priority areas are 1) The capacity and limitations of law in affecting the behavior of individuals and organizations, the conditions that enhance or diminish the impact of law, and the processes by which this impact is achieved or undermined 2) The personal, cultural, and social factors that affect the use of law and law-like systems in dispute settlement, including factors that limit access to or knowledge about the relevant legal processes 3) The operation of informal systems, including negotiation and arbitration, for processing legally relevant disputes. 4) The causes and processes of change in legal systems. Research that promises to explore methods of potentially wide utility in the social scientific study of law through the development of new techniques or the application of existing techniques in innovative ways will be given priority in the competition for funding. For further information, write: Program Director, Law and Social Sciences Program, Division of Social Sciences, National Science Foundation, Washington, D.C. 20550.

The Universities-National Bureau Committee for Economic Research invites applications for membership by interested universities. The Committee holds a meeting once a year on a topic selected by the members. The conference proceedings which now include thirty volumes have been published as a series by the National Bureau of Economic Research since 1949. Universities interested in membership should apply to the Chairman of the Committee: Professor Edwin S. Mills, Department of Economics, Princeton University, Princeton, NJ 08540. The criterion for acceptance of a university as a member of the Committee is the extent and quality of economic research carried on at the university.

The fifth Annual History of Economics Conference will be held at the University of Toronto, May 25-27, 1978. Persons wishing to present papers are invited to submit abstracts (2 copies) together with *separate* sheets containing: Author's name, address, professional affiliation, position, and telephone number; title of paper; whether or not the author is a member of the Society. These are to be received no later than October 1, 1977. All communications to be sent to Crauford D. Goodwin, Department of Economics, Duke University, Durham, NC 27706.

The meeting of the Gesellschaft für Wirtschafts und Sozialwissenschaften (Verein für Socialpolitik) will be held September 25-27, 1978 in Hamburg. The theme will be "The State and the Economy." A program committee, Professors P. Eichhorn (Speyer), G. Gäfgen (Konstanz), C. Seidl (Graz), and C. C. v. Weizsäcker (Bonn), will select the papers to be presented. Applicants are to send their proposals for papers to the chairman: Professor Dr. C. C v. Weizsäcker, Department of Economics, University of Bonn, Adenauerallee 24, D-5300 Bonn.

Deaths

Ira B. Cross, Menlo Park, California, Mar. 24, 1977.

John A. Delehanty, associate professor, department of economics, Kansas State University, Apr. 4, 1977.

Harry G. Johnson, University of Chicago and London School of Economics, May 9, 1977.

Egon Sohmen, professor of economics, University of Heidelberg, Mar. 8, 1977.

Retirements

Raymond J. Doll, senior vice president and director of research, Federal Reserve Bank of Kansas City, Feb. 28, 1977.

Visiting Foreign Scholars

Jack Johnston, University of Manchester, England: visiting professor, Emory University, Sept. 1, 1977.

Promotions

Tridib Kumar Biswas: senior economist, Ministry of Transportation and Communications, Toronto, May 2, 1977.

Harvey Botwin: professor of economics, Pitzer College-Associated Colleges at Claremont, Sept. 1977.

Leon E. Danielson: associate professor of economics and business, North Carolina State University, July 1, 1977.

Thomas E. Davis: senior vice president and director of research, Federal Reserve Bank of Kansas City, Mar. 1, 1977.

Edward J. Frydl: chief, Balance of Payments Division, Federal Reserve Bank of New York, Feb. 17, 1977.

David E. R. Gay: associate professor of economics, University of Arkansas, July 1977.

Henry A. Gemery: professor of economics, Colby College

Edgar Walton Jones: professor of economics and business, North Carolina State University, July 1, 1977.

Thomas R. McKinnon: associate professor of economics, University of Arkansas, July 1977.

James W. Meehan, Jr.. associate professor of economics, Colby College.

Lloyd D. Orr: professor of economics, Indiana University, July 1, 1977.

James R. Peeler, Jr. professor of economics and business, North Carolina State University, July 1, 1977.

Richard K. Perrin: professor of economics and business, North Carolina State University, July 1, 1977.

Rutbert D. Reisch: chief, Industrial Economies Division, Federal Reserve Bank of New York, Mar. 31, 1977.

Charles R. Roll: economist, economics department, The Rand Corporation, Feb. 1977.

Richard E. Sylla, professor of economics and business, North Carolina State University, July 1, 1977.

Administrative Appointments

Ryan C. Amacher: chairman, department of economics, Arizona State University, Aug. 1977.

Martin S. Feldstein: president, National Bureau of Economic Research, Apr. 22, 1977.

William F. Ford: senior vice president and chief economist, Wells Fargo Bank, San Francisco, Mar 1977.

A. Ray Grimes, Jr.. chief, Regional Economic Analysis Division, Bureau of Economic Analysis, U S Department of Commerce, Mar. 28, 1977.

W. Whitney Hicks: chairman, department of economics, University of Missouri-Columbia, Jan. 1, 1977.

Martin T. Katzman, Harvard University: head and professor, program in political economy, University of Texas-Dallas, Sept. 1, 1977.

Charles E. McLure, Jr.: executive director for research, National Bureau of Economic Research, Apr. 22, 1977.

Gerard E. Nistal, Southeastern Louisiana Univer-

ity: chairman, division of business and economics,)ur Lady of Holy Cross College, Jan. 1977.

James J. O'Leary: chairman, National Bureau of :conomic Research, Apr. 22, 1977.

New Appointments

Steven L. Balch: assistant economist, economics de-artment, The Rand Corporation, Jan. 1977.

Stephen P. A. Brown: assistant professor, depart-ıent of economics, Arizona State University, Aug. 977.

Winston Chow: associate statistician, economics de-ıartment, The Rand Corporation, Oct. 1976.

Hope Corman: instructor, department of economics, lutgers-The State University, July 1, 1977.

George C. Eads, National Commission on Supplies .nd Shortages: senior economist, economics depart-nent, The Rand Corporation, May 1977.

Evelyn Fallek: economist, Banking Studies Division, 'ederal Reserve Bank of New York, Aug. 3, 1976.

Benjamin M. Friedman: research associate, Na-ıonal Bureau of Economic Research.

Carl M. Gambs: financial economist, Federal Re-ıerve Bank of Kansas City, Aug. 1976.

Nancy A. Garvey: instructor, department of eco-ıomics, Rutgers-The State University, July 1, 1977.

Malcolm Getz. research associate, National Bureau of Economic Research, June 1977.

Gary A. Gigliotti instructor, department of eco-nomics, Rutgers-The State University, July 1, 1977.

Thomas K. Glennan, Jr., National Academy of Sciences/National Research Council: senior econ-omist, The Rand Corporation, Washington, D C., Sept 1976.

William W. Gould. research associate, National Bureau of Economic Research, Sept. 1977.

John D. Hanson. assistant professor, department of economics, North Carolina State University, Apr. 1, 1977

Bryon M. Higgins: financial economist, Federal Re-serve Bank of Kansas City, Oct. 1976.

Michael D. Hurd: visiting research fellow, National Bureau of Economic Research, Sept. 1977.

Bruce Kaufman, University of Wisconsin: assistant professor of economics, Georgia State University, Sept. 1977

Michael Kennedy, University of Texas-Austin associate economist, The Rand Corporation, June 1977.

Lawrence W. Kenny: visiting scholar, National Bureau of Economic Research, Sept. 1977.

Elsie M. Knoer: assistant professor, department of economics, Arizona State University, Aug. 1977.

C. Timothy Koeller: instructor, department of eco-nomics, Rutgers-The State University, July 1, 1977.

Walter C. Labys: research associate, National Bureau of Economic Research.

Lee Lillard, National Bureau of Economic Re-search economist, economics department, The Rand Corporation, Sept. 1977.

Barbara R. McIntosh: instructor, department of economics, Rutgers-The State University, July 1, 1977.

Lee R, McPheters, Florida Atlantic University: associate professor, department of economics, Arizona State University, Aug. 1977.

Jorge Martinez, Washington University: assistant professor of economics, Georgia State University, Sept. 1977.

Mathew J. Morey: assistant professor, department of economics, Arizona State University, Aug. 1977.

Winford C. Naylor, The Brookings Institution: assistant professor of economics, Pitzer College-Associated Colleges at Claremont, Sept. 1977.

Mieko Nishimizu: visiting research fellow, National Bureau of Economic Research, Sept. 1977.

Edgar Ortiz: instructor, department of economics, Rutgers-The State University, July 1, 1977.

James L. Pierce: research associate, National Bureau of Economic Research.

Maury R. Randall: assistant professor, department of economics, Rutgers-The State University, July 1, 1977.

Jeffrey S. Royer: assistant professor, department of economics and business, North Carolina State University, July 1, 1977.

Daniel A. Seiver. visiting research fellow, National Bureau of Economic Research, Sept. 1977.

Richard H. Thaler: research associate, National Bureau of Economic Research.

Thomas Tietenberg: associate professor, depart-ment of economics, Colby College.

Richard B. Victor, University of Michigan: asso-ciate economist, The Rand Corporation, Washington, D.C., Apr 1977.

Gary F. Vocke: assistant professor, department of economics and business, North Carolina State Univer-sity, July 1, 1977.

Wayne Vroman: research associate, National Bu-reau of Economic Research, Apr. 22, 1977.

Steven Weisbrod: economist, Banking Studies Divi-sion, Federal Reserve Bank of New York, Aug. 17, 1977.

Michael K. Wohlgenant: assistant professor, depart-ment of economics and business, North Carolina State University, July 1, 1977.

Leaves for Special Appointments

Todd Sandler, University of Wyoming: NATO fel-low, Institute of Social and Economic Research, Uni-versity of York, England, June 1, 1977-May 31, 1978.

Daniel A. Seiver, ISER, University of Alaska: Na-tional Bureau of Economic Research faculty research fellow, Cambridge, Mass., 1977-78.

Simon Teitel, Catholic University of America: advi-sor, department of economics and social development, Inter-American Development Bank.

Resignations

Dana N. Stevens, Georgia State University: New College, University of Southern Florida, Sept., 1977.

Miscellaneous

H. Michael Mann, Boston College: Managing Edi-tor/USA, *Journal of Industrial Economics*, May 15, 1977.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment his new title (if any), and the date at which the change will occur.

C. Type each item on a separte 3×5 card and please do not send public relations releases

D. The closing dates for each issue are as follows. *March*, November 1; *June*, February 1; *September*, May 1 *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry

By WILLIAM J. BAUMOL*

The literature is generally somewhat vague on the circumstances in which monopoly offers cost advantage over production by a multiplicity of firms. The sense of the discussion seems to be that monopoly will be "natural,"[1] that is, it can provide outputs at lower social costs, when and only when the industry has economies of scale. Perhaps the most unexpected finding of this paper is that *scale economies are neither necessary nor sufficient for monopoly to be the least costly form of productive organization.* Rather, the critical concept is (by definition) strict subadditivity of the cost function, meaning that the cost of the sum of any *m* output vectors is less than the sum of the costs of producing them separately. Since subadditivity is a mathematical concept whose properties do not seem to have been explored fully, a good part of our task will be the characterization of subadditive cost functions and their relation to more conventional concepts such as declining average costs (which, as we will see, it is inaccurate to equate to scale economies).

I

My discussion will begin with the single product firm. For these we will show that evidence of scale economies is always *sufficient* but *not necessary* to prove subadditivity. That is, it is (much) too demanding a test of natural monopoly to require evidence of scale economies or declining average costs, *even in the neighborhood of current output levels.* Nevertheless, if anything, our analysis makes it harder to prove that a particular monopoly is natural, even in the single product case. For it turns out that proof of subadditivity requires a *global* description of the shape of the *entire* cost function from the origin up to the output in question, thus calling for data that may lie well beyond the range of recorded experience.

When we turn to the multiproduct case we will find that sufficient conditions for subadditivity must include some sort of complementarity in the production of the different outputs of the industry which corresponds to a type of convexity to be specified in the next section.

Our discussion will proceed on the assumption that the menu of available techniques is fixed (no technological change), that exactly the same menu of techniques is available to the monopolist and to each of its potential competitors, *and that all input prices are fixed.*[2]

*Professor of economics, Princeton and New York universities. This paper was prepared under the sponsorship of the Division of Science Information of the National Science Foundation as part of a study of scale economies and public good elements in information transfer. The writing of this paper was very much a group undertaking with the most critical contributions made by Elizabeth Bailey and Dietrich Fischer. Important suggestions were also made by M.I. Nadiri, Janusz Ordover, Thijs ten Raa, Michael Rothschild, and Robert Willig as the analysis gradually evolved.

[1]For an interesting discussion of the origins of the concept and the term see Edward Lowry.

[2]Since we assume all input prices, w, to be constant, the cost function $C(y, w)$ can be written simply as $C(y)$, where y is the vector of outputs. Input prices are taken to be constant because an industry may be a natural monopoly at one set of input prices and not at another. For example, where capital is cheap relative to labor, single firm production may be cheapest, while the reverse may be true where low wages call for production techniques that are not capital intensive. Therefore, it seems to me that the appropriate test to determine whether an industry is *now* a natural monopoly must indicate whether it is the least costly market form at (or in the neighborhood of) current input prices. To require this to be true for all input prices would be far too restrictive.

II. Definitions[3]

A rather tedious section on definitions is unavoidable because some of the concepts have not been used widely in the literature and the extension of others to the n-product case is not entirely straightforward.

There seems to be some ambiguity in the term "natural monopoly" which is used to refer to one or both of two circumstances:[4]

a) An industry in which multifirm production is more costly than production by a monopoly (subadditivity of the cost function).

b) An industry to which entrants are *not* "naturally" attracted, and are incapable of survival even in the absence of "predatory" measures by the monopolist (sustainability of monopoly).

This article deals exclusively with the first of these concepts, leaving the issue of sustainability to other papers. (See Faulhaber 1975a; Baumol, Bailey, and Willig; John Panzar and Willig 1975b.)

Accordingly, we begin our formal definitions with our basic criterion of natural monopoly which is given by

DEFINITION 1: *Strict and Global Subadditivity of Costs.* A cost function $C(y)$ is strictly and globally subadditive in the set of commodities $N = 1, \ldots, n$, if for any m output vectors $y^1, \ldots, y^m$ of the goods in N we have

$$C(y^1 + \ldots + y^m) < C(y^1) + \ldots + C(y^m)$$

This is clearly the necessary and sufficient condition for natural monopoly of any output combination in the industry producing (any and all) commodities in N, for subadditivity means that it is always cheaper to have a single firm produce whatever combination of outputs is supplied to the market, and conversely.

Of course, it is possible that for some output vectors an industry will be a natural monopoly while for others it will not. In such a case we have output-specific subadditivity, meaning that the pertinent output vector y^* is produced more cheaply by one firm than by any combination of smaller firms.

We must also define precisely the other pertinent cost attributes, all of which acquire somewhat novel features in an n-product firm or industry. We begin with

DEFINITION 2: *Strict Economies of Scale* in the production of outputs in N are present if for any initial input-output vector $(x_1, \ldots, x_r, y_1, \ldots, y_n)$ and for any $w > 1$, there is a feasible input-output vector[5] $(wx_1, \ldots, wx_r, v_1 y_1, \ldots, v_n y_n)$ where all $v_i \geq w + \delta, \delta > 0$.

This definition in effect tests for scale economies by considering any w-fold expansion in all input quantities, and requiring that it permit *at least* a $w + \delta$-fold proportionate increase in each output. Note that if the production function is not homothetic *the firm may not wish to expand its input usage proportionately*, and depending on demand relationships, it certainly may not want to increase its *outputs* proportionately. Thus, our criterion of scale economies is expressed entirely in terms of hypothetical increases in input and output quantities along rays in input and output spaces, respectively.

Declining average costs, a concept usually associated with scale economies, is not so readily extended to a multiplicity of outputs. The problem, of course, is that if outputs do *not* expand proportionately we do not know how to define an index of aggregate output by which to divide total cost, nor do we have any way of apportioning

[3]For a fundamental article covering ground related to much of the substance of this paper see Peter Newman. For illuminating discussions of some of the definitional problems in this area and some related matters see also Giora Hanoch (1970, 1975) and also W. W. Sharkey and L. G. Telser.

[4]See, for example, Richard Posner. Gerald Faulhaber (1975a) has shown with the aid of ingenious and illuminating counterexamples that the two conditions are not equivalent; specifically that condition a) does not imply satisfaction of condition b). Baumol, Elizabeth Bailey, and Robert Willig have demonstrated the converse—that b) does imply a).

[5]I write v_i rather than v to allow for the case where *proportionate* expansion of outputs is not possible. If we assume free disposal then there is no need to do so, since if the percentage increases in some outputs were to exceed $w + \delta$ we could simply get rid of the excess.

joint and common costs so as to calculate an average cost, item by item.[6] For our purposes, however, it will suffice to deal only with the special case in which output quantities all happen to vary proportionately *but input quantities follow the least-cost expansion path,* which in general does not involve proportionate changes in input quantities. Accordingly, we formulate

DEFINITION 3: *Ray Average Costs (RAC)* are *strictly declining* when

$$(1)\quad C(vy_1, \dots, vy_n)/v < C(wy_1, \dots, wy_n)/w \text{ for } v > w$$

where v and w are measures of the scale of output along the ray through

$$y = (y_1, \dots, y_n)$$

As Definitions 2 and 3 clearly show, scale economies and declining average costs in the sense we have used it, are both ray-specific properties. We will use one other ray-specific property:

DEFINITION 4: *Ray Concavity.* A cost function $C(y)$ is strictly output-ray concave (marginal costs of output *bundles* everywhere decreasing) if

$$(2)\quad C[kv_1y + (1-k)v_2y] > kC(v_1y) + (1-k)C(v_2y)$$

for any

$$v_1, v_2 \geq 0, \qquad v_1 \neq v_2, \qquad 0 < k < 1$$

This definition can be extended in an obvious way to m rather than two output vectors. The parameters v_1 and v_2 simply assure us that v_1y and v_2y are two output vectors that lie on the same ray (their outputs are proportional). The parameter k plays its usual role in the standard definition of concavity.

In the n-output case we must also have a way of characterizing the behavior of costs as output *proportions* vary. One pattern of such "transray" behavior will prove critical for our analysis. This concept is perhaps best described as a new formal interpretation of the phenomenon of complementarity in the production of the various goods and services supplied by the firm. We call a cost function *transray convex* at output vector $y = (y_1, \dots, y_n)$ if along at least one hyperplane through y, $\Sigma w_i y_i = w$, a weighted average of the costs of producing separately any two output vectors on this hyperplane is no less than the cost of producing any weighted average of those two outputs *together.* That is, in geometric terms, a total cost function is transray convex at y if along some negatively sloping cross section through y in output space, costs are lower (no higher) toward the interior of the cross section than they are toward its edges (curve $C^*C^{*\prime}$ in Figure 4b), so that it is no more expensive to produce goods in combination rather than separately. More formally:

DEFINITION 5: *Transray Convexity.* A cost function $C(y)$ will be called transray convex through $y^* = (y_1^*, \dots, y_n^*)$ if there exists any set of positive constants $w_1, \dots, w_n$, such that for every two output vectors $y^a = (y_{1a}, \dots, y_{na})$, $y^b = (y_{1b}, \dots, y_{nb})$ lying in the same hyperplane $\Sigma w_i y_i$ through y^*, that is, for which $\sum_i w_i y_{ia} = \sum_i w_i y_{ib} = \sum_i w_i y_i^*$, we have

$$(3)\quad C[ky^a + (1-k)y^b] \leq kC(y^a) + (1-k)C(y^b)$$

for any k, $\quad 0 < k < 1$

This concept is closely related to what Panzar and Willig have named "economies of scope," in contradistinction to economies of scale.

Together, the concepts of strictly declining ray average costs and that of transray convexity turn out to be an extremely powerful combination offering us a new characterization of a cost function that is well behaved for a number of analytical purposes. We will see later (Proposition 12) that they are sufficient for subadditivity of costs. Moreover, in a closely related paper (Baumol, Bailey, and Willig), it is shown

[6]However, there is a natural extension of the concept of declining average *variable* cost of one individual product to the multioutput case. See the author (1976).

that these go far toward giving us conditions sufficient for the existence of some monopoly prices that are sustainable against entry. It may be suspected that they will prove to have many other analytical implications.

III. Some Interrelationships Among the Cost Concepts

Having completed the tedious task of definition of our concepts, I can now summarize their relationships briefly, indicating which of them implies which. This is done in outline in Figure 1 which shows that global scale economies and decreasing ray average costs imply subadditivity along a ray, but that the former are not necessary for the latter. Moreover, they are neither sufficient nor necessary either for strict and global subadditivity, or for strict and output-specific subadditivity, the requirements for natural monopoly in an n-product industry.

We now proceed to prove some of the propositions underlying Figure 1. Before turning to the central issue of subadditivity we must explore briefly a few relationships among scale economies, ray concavity, and declining ray average costs. Among these cost concepts, scale economies occupies a position by itself, because all the others assume that inputs adjust optimally to output changes, that is, inputs are taken to follow along the expansion path of least costs. On the other hand, the concept of scale economies requires inputs to change proportionately, and in general, this will not minimize the cost of an expansion. (See, for example, Hanoch, 1975.) Thus, if unit cost decreases when inputs are increased *proportionately*, they must certainly decrease along the least cost expansion path. This should at least suggest the logic of a result which I have proven elsewhere (1976):

PROPOSITION 1: *Strict economies of scale are sufficient but not necessary for ray aver-*

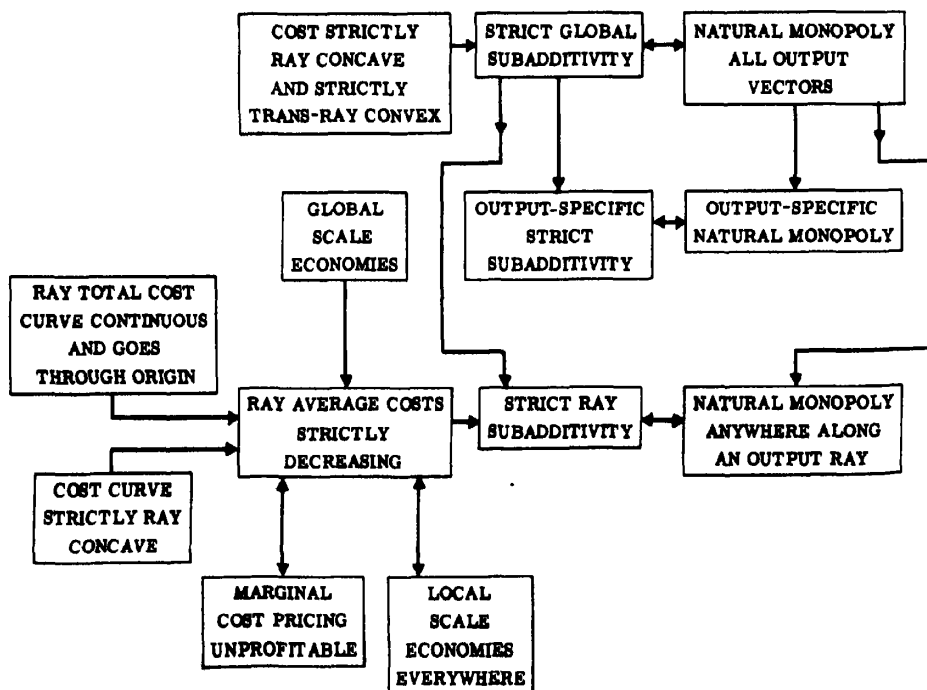



FIGURE 1

age cost to be strictly declining.

We also have a relation between strict concavity and ray average costs:

PROPOSITION 2: *Strict ray concavity (declining marginal cost) of the cost function $C(y)$ and $C(0) = 0$ after subtraction of any fixed costs, are sufficient but not necessary for ray average costs to decline.*

PROOF:
By the definition of strict ray concavity we have by (2) for $v_1 = 1, v_2 = 0$,

$$(4) \quad C(ky) > kC(y) \quad \text{for} \quad 0 < k < 1$$

which is criterion (1) of declining ray average cost.

Since it is well known that average costs can fall even when marginal costs are rising (nonconcavity in total cost), the result follows.

IV. Ray Subadditivity and its Relation to Other Cost Attributes[7]

Since subadditivity is the defining attribute of a natural monopoly, the heart of our task consists in the analysis of this property and its relation to the other cost attributes. The next few sections deal with scale behavior—subadditivity along a ray—leaving the true multiproduct case until later. First we prove

PROPOSITION 3: *Strictly declining ray average cost implies strict ray subadditivity.*

PROOF:
Consider the n output vectors $v_1y, \ldots, v_ny$ along the same ray, all $v_i > 0$. By the definition of ray average costs that are strictly declining

$$(5) \quad C(\Sigma v_j y)/\Sigma v_j < C(v_i y)/v_i \qquad (i = 1, \ldots, n)$$

Consequently, $C(\Sigma v_j y)/\Sigma v_j$ is less than a weighted average of the $C(v_i y)/v_i$, i.e.,

$$\frac{C(\Sigma v_j y)}{\Sigma v_j} < \sum \frac{v_i}{\Sigma v_j} \frac{C(v_i y)}{v_i} = \frac{\Sigma C(v_j y)}{\Sigma v_j}$$

which immediately yields our subadditivity result

$$C(\Sigma v_j y) < \Sigma C(v_j y)$$

It also follows from Propositions 2 and 3 that strict ray concavity, along with $C(0) = 0$, is sufficient for strict ray subadditivity. However, as I will show now, the converse does not hold—ray subadditivity is *not* sufficient either for ray concavity, or for declining ray average cost.

It is easy to produce the necessary counterexamples to prove this negative result—cost functions that are strictly ray subadditive and yet not ray concave throughout, and for which ray average cost does not decline throughout. An extreme case is shown in Figure 2a—the piecewise-linear cost function $OABDC$. This is clearly *not* concave, as is shown by the portion of the cost curve which lies below line segment WS. Similarly, average cost *increases* along part of the cost curve: output y_s is greater than y_r, yet the slope of ray OS is greater than the slope of OR so that ray average cost at y_s is greater than ray average cost at y_r.

To complete the argument I must now show that the cost function is strictly subadditive. A moment's consideration shows why this is so. In the case depicted, total cost for any output produced by one firm alone cannot exceed $OA + BD$. Similarly, total cost of any output produced by two or more firms in any way cannot be less than $2(OA)$, the fixed costs of two firms together. Since in our cost function $OA > BD$, we have: total cost of any output by a single firm $\leq OA + BD < 2(OA) \leq$ total cost of production by more than one firm.

We have thus proved by counterexample our desired result:

PROPOSITION 4: *Neither ray concavity nor ray average costs that decline everywhere are necessary for strict subadditivity.*

We may note also that while the extreme case shown in Figure 2a makes the proof

[7] A number of results in this section were first derived by Faulhaber (1975b).

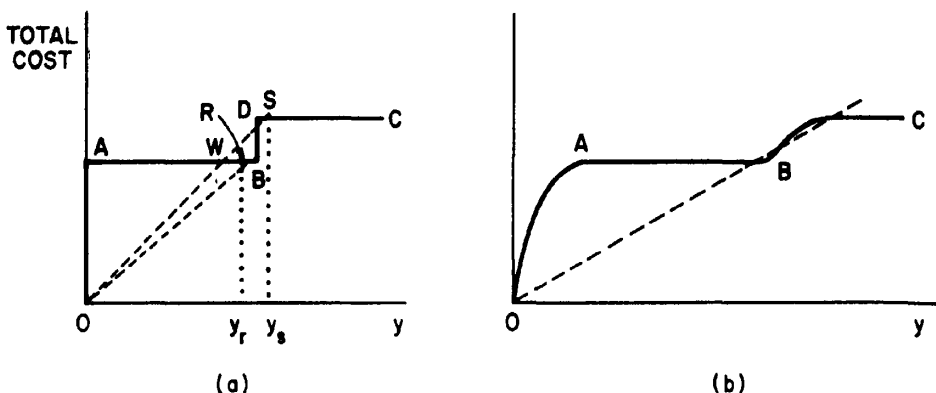



FIGURE 2

easier, less restrictive counterexamples are easy to supply. By including a series of steps in the graph of the cost function instead of the single step of Figure 2a, one can clearly produce a subadditive cost function in which average costs are rising at a number of intervals. Moreover, as Figure 2b illustrates, intervals of rising ray average costs are compatible with a completely smooth cost function that is strictly subadditive.

From Propositions 3 and 4, and since by Proposition 1 economies of scale are sufficient for decreasing ray average costs, we immediately have the following key result:

PROPOSITION 5: *Global scale economies are sufficient but not necessary for (strict) ray subadditivity, the condition for natural monopoly in the production of a single product or in any bundle of outputs produced in fixed proportions.*

V. On Characterization of Ray Subadditivity

Before going further it is appropriate to say something about the characteristics of subadditive (cost) curves. We have seen that they need not be concave, nor need they exhibit declining average costs. What then do they look like in general? Three necessary conditions for subadditivity in a two-dimensional graph may offer the reader some of the pertinent flavor. First we have

PROPOSITION 6: *Strict subadditivity along a ray implies that for any output vector, y, in that ray, $C(y) < vC(y/v)$ for v any integer ≥ 2. That is, for any such integer v, ray average cost must on the average decrease over the interval between y/v and y.*

This result follows from the definition of subadditivity which requires, for example, for $v = 3$, that $C(y) < C(y/3) + C(y/3) + C(y/3) = 3C(y/3)$ and a similar expression obviously holds for any other integer value of v. Thus, in Figure 3a the height of our cost curve at output y^* is C^*. This gives us $C^*/2$, $C^*/3$, $C^*/4$, etc., as respective *floors* for total costs for outputs $y^*/2$, $y^*/3$, $y^*/4$, etc., as indicated by the heights of the vertical line segments. If the cost function is to be subadditive, then at output level $y^*/2$ total cost must be represented by some point such as D which is *above* $C^*/2$, and the analogous observation applies to points B and A.

PROPOSITION 7: *If a function $C(y)$ is strictly subadditive, nonnegative and increasing in value along a ray, then either $c(0) > 0$ or the function must be strictly concave for y in the neighborhood of the origin.*

For suppose $C(0) = 0$, writing $k = 1/v < 1$ in Proposition 6 we have $kC(y) < C(ky)$ in the neighborhood of $k = 0$. But $kC(y) \equiv (1 - k)C(0) + kC(y)$ and $C(ky) \equiv C[(1 - k)0 + ky]$. Consequently, in the limit as k approaches zero we must

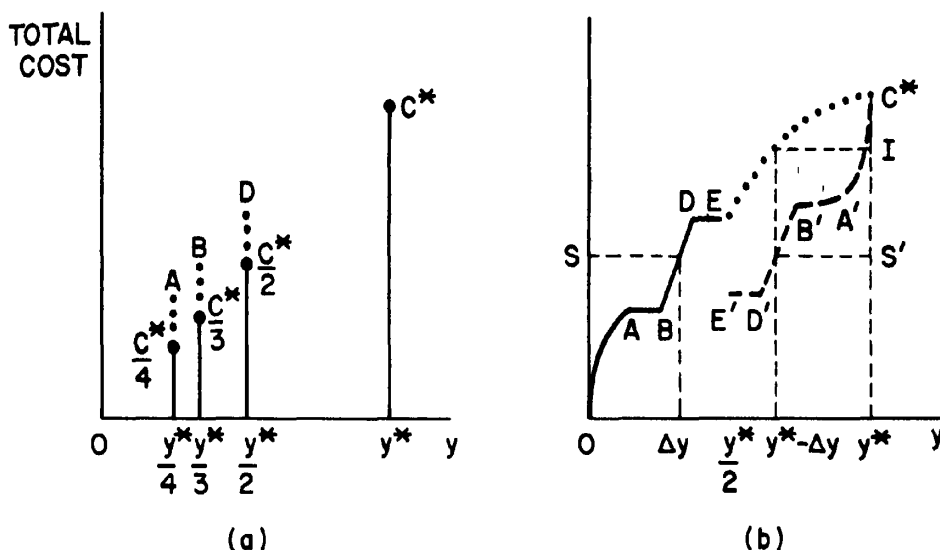



FIGURE 3

satisfy the requirement of strict concavity

$$(1 - k)C(0) + kC(y) < C[(1 - k)0 + ky]$$

A third necessary condition for subadditivity is given by

PROPOSITION 8 (Dietrich Fischer): *If $C(y)$ is strictly subadditive along a ray and y^* is any particular output vector on that ray, with $C(y^*) = C^*$, then the graph of $C(y^* - \Delta y)$ in the interval $y^*/2 \leq y^* - \Delta y \leq y^*$ must everywhere lie above the inverted*[8] *mirror image of $C(\Delta y)$ in the interval $0 \leq \Delta y \leq y^*/2$. That is, since by subadditivity $C(y^* - \Delta y) + C(\Delta y) > C(y^*)$, then*

$$(6) \quad C(y^* - \Delta y) > C(y^*) - C(\Delta y)$$

As (6) shows, Proposition 8 follows directly from the definition of subadditivity. Condition (6) also tells us, incidentally, that with subadditivity the increment in total cost resulting from an addition Δy to an output, must be less than the cost $C(\Delta y)$ of producing that addition all by itself (its "stand-alone cost").

Figure 3b indicates more explicitly the graphic implications of Proposition 8. Let the total cost curve in the interval between $y = 0$ and $y = y^*/2$ be $OABDE$. Now starting from C^*, the cost corresponding to y^*, draw (backwards) the curve $E'D'B'A'C^*$ which is the "upside-down" mirror image of $OABDE$. For example, with concave arc OA curving upward and to the right from O, arc C^*A' will be convex and curve downward and to the left from C^*. Now, Proposition 8 requires that the actual cost curve over the interval between $y^*/2$ and y^* (dotted curve EC^*) must everywhere lie above $E'D'B'A'C^*$.

This requirement is designed to assure us that the "incremental cost" IC^* of adding output Δy to output $y^* - \Delta y$ is less than the stand-alone cost OS of producing output Δy by itself, since, by construction, $OS = S'C^* > IC^*$.

VI. Some Implications for Evidence on Natural Monopoly

Propositions 6, 7, and 8 show that to test whether subadditivity is satisfied *just at a particular output level*, y^*, we can *not* examine only the cost curve in the vicinity of y^*. Because a claim of natural monopoly asserts that production by a single firm is cheaper than it would be in the hands of

[8]The center of inversion is the point $(y^*/2, C^*/2)$.

any and every possible combination of smaller firms, we must know the behavior of the cost curve *throughout its length in the interval between the origin and* y^*. This is a rather difficult prospect for those who bear the burden of proof of the presence or absence of natural monopoly!

On the other hand, Propositions 4 and 5 imply that the tests sometimes suggested to evaluate the cost advantages or disadvantages of monopoly have in another respect been excessively demanding. In a number of regulatory hearings, much stress has been placed on the allegation that the regulated firm in question had reached a size where "economies of scale have been exhausted," implying that since output is now being produced under constant ray average and perhaps marginal costs, several firms can provide it as cheaply as one. But this is simply not true in general (on this see, for example, Seneca). Indeed, it is easy to see that single firm production will always be cheaper so long as average costs decline strictly monotonically up to any output $(y^*/2) + \Delta$ for any $\Delta > 0$ and are level thereafter, where y^* is total industry production. For if two more firms were to provide the commodity instead, at least one such firm must produce no more than half of industry output and must therefore incur average costs greater than the minimum average cost attainable. That is,

PROPOSITION 9: *A condition sufficient for strict and output-specific subadditivity along a ray is that, with* y^* *the given output vector of the industry, ray average cost declines strictly throughout the interval between zero and* $y^*/2 + \Delta$ *for some* $\Delta > 0$, *and that it declines or remains constant in the interval to the right of* $y^*/2 + \Delta$ *and through* y^*.

VII. Transray Behavior and Multiproduct Subadditivity

From the single product case, we turn finally to the conditions under which a single firm is the most efficient supplier of all the products provided by some industry or at least of some specified subset of those outputs.

Since scale economies refer only to the technical gains from increases in volume of output rather than those accruing from production of several goods together, it is obvious that by themselves they cannot tell us whether it is cheaper to produce two different output bundles separately or together. The issue is one relating to the behavior of the cost function from one output ray to another, rather than along any one ray. It is therefore naturally to be suspected that evidence of the presence of scale economies will be insufficient to guarantee what we may refer to as transray subadditivity.

If that were all there is to the matter, the insufficiency of scale economies for natural monopoly might seem a trivial and uninteresting observation. But there is more to the issue.

The obvious extension of the concept of scale economies to the entire cost function is concavity. Concavity is, after all, a rather strong assumption. But it is easy to show that even strict concavity is neither sufficient nor necessary for subadditivity in an n-product cost function. First we show:

PROPOSITION 10: *Strict concavity of a cost function is not sufficient to guarantee subadditivity.*

PROOF by counterexample:[9]

Consider the cost function in two outputs

$$C = y_1^a + y_1^k y_2^k + y_2^a \qquad 0 < a < 1, \quad 0 < k < 1/2$$

Since the sum of several concave functions is itself concave, it is clear that the cost function is strictly concave because of the values of k and a. However, we have for

[9] I am indebted for this counterexample to Dietrich Fischer. It may seem at first that a case of superadditivity such as that in the example cannot occur in reality since the firm which finds it more expensive to produce two items together will turn them out separately (in different plants). But administrative and communication costs can still make it more expensive to produce this way than in two totally independent firms. That is why, in reality, some industries are characterized by many small firms, each with their different specializations. The cost functions are such that the giant multiproduct firm just cannot compete.

example, for $y_1^* = y_2^* = 1$:

$$C(y_1^*, 0) = 1 \qquad C(0, y_2^*) = 1$$
$$C(y_1^*, y_2^*) = 3 > C(y_1^*, 0) + C(0, y_2^*) = 2$$

Hence the function is not subadditive.

We also have immediately the basic result

PROPOSITION 11: *Scale economies are neither necessary nor sufficient for subadditivity.*

Nonnecessity was already proved in Proposition 5. To show insufficiency we need merely take the production function implicit in the preceding example to involve a single input (whose quantity is x) and whose price p is fixed. Then substituting $C = px$ into the cost function we see that it obviously exhibits scale economies throughout even though it is not subadditive.

It must be emphasized that this example does not represent some sort of pathological exception. Rather, as we will see next, strict concavity of a cost function involves an attribute which works in the direction opposite to that called for to yield transray subadditivity.

The main issue is what sorts of departure from concavity favor subadditivity? The answer is suggested by Figures 4a and 4b. Figure 4a depicts a cost function represented by surface OCC' which is strictly concave and exhibits scale economies throughout. Three cross sections along rays Ra, Rb, and Rc show the characteristic concave shape and, hence, the declining average costs which we have seen to suffice for subadditivity along a ray.

However, despite the concavity *and the presence of scale economies*, the cross section $CABDC'$ taken across these rays tells quite a different story. Because its lowest points are reached at the y_1 and y_2 axes, this concave transray cross section favors production of commodities in isolation. That is, it makes for increased cost of production using common facilities. This is precisely the opposite of what is wanted for transray subadditivity (or for an *interior* cost minimum along a fixed iso-revenue curve, as is indicated by Figure 4c, the iso-cost map corresponding to Figure 4a). Rather, subadditivity is favored by a cost function which, because of complementarity in production, is shaped like that in Figure 4b whose transray cross section $C^*A^*B^*D^*C^{*\prime}$ reaches its lowest points in the interior of the diagram where both commodities are produced together. The corresponding iso-cost map (Figure 4d) is obviously also conducive to an interior cost minimum along an iso-revenue locus.

We come now to our key result:

PROPOSITION 12: *Ray average costs that are strictly declining and (nonstrict) transray convexity along any one hyperplane $\Sigma w_i y_i = w$, $w_i > 0$ through an output vector y are sufficient to guarantee strict output-specific subadditivity for output y.*

PROOF:

The result has already been proved in Proposition 3 for the division of y into any y^a, y^b which lie in the same ray. It therefore only remains to prove for any $y^a + y^b = y$

$$C(y^a + y^b) < C(y^a) + C(y^b) \text{ if } y^a \neq wy^b$$

By the definition of ray average costs that are strictly declining we have for any $v > 1$

$$vC(y) > C(vy) \tag{7}$$

Let $w_1, \ldots, w_n$ be any vector of positive constants defining the cross section along which y satisfies Definition 5 of transray convexity. Now define[10] v_a, v_b by $v_a \Sigma w_i y_{ia} = \Sigma w_i y_{ia} + \Sigma w_i y_{ib} = v_b \Sigma w_i y_{ib}$ and set

$$k = 1/v_a \tag{8}$$

so that

$$1/v_b = 1 - 1/v_a = (1 - k)$$

[10]As is illustrated in Figure 5, the purpose of multiplying all the elements of y^a and y^b by v_a and v_b, respectively, is to expand each of these outputs along its own ray until it is on the same hyperplane of transray convexity $\Sigma w_i y_i = w$, which contains point $y^a + y^b$. That then permits us to apply Definition 5 of strict transray convexity to the three points $v_a y^a$, $v_b y^b$, and $y^a + y^b$, to show that the cost of the latter does not exceed a weighted average of the costs of the former.

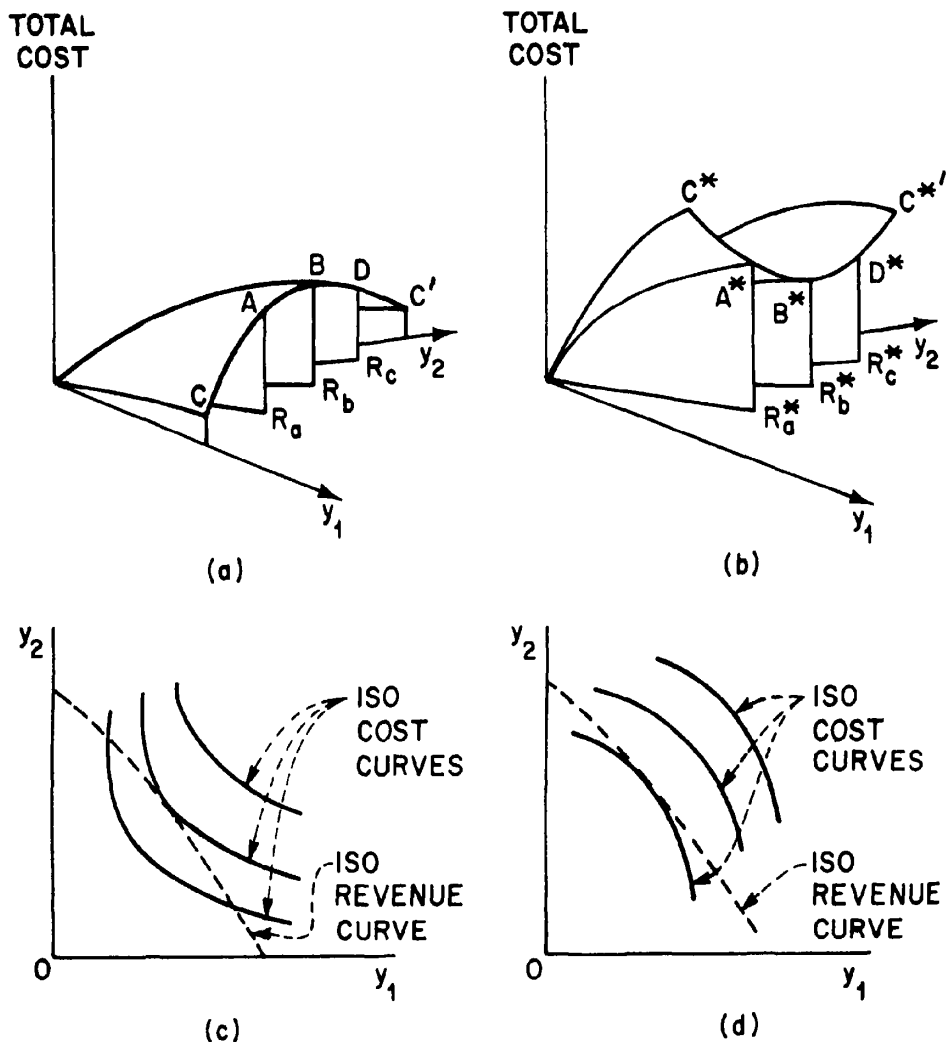



FIGURE 4

then by (7) and (3) in Definition 5

$$kv_aC(y^a) + (1 - k)v_bC(y^b) > kC(v_ay^a) + (1 - k)C(v_by^b) \geq C[kv_ay^a + (1 - k)v_by^b]$$

or since $kv_a = (1 - k)v_b = 1$ by (8) we have our subadditivity result

$$C(y^a) + C(y^b) > C(y^a + y^b)$$

Note that the proof does not require *strict* transray convexity, but it does have to assume that ray average costs are *strictly* declining over *part* of the region since the proof of Proposition 3, which is subsumed here, does require the latter. If, along the ray containing the industry output vector, average costs were not strictly declining, for example, if they increased proportionately with output, then several smaller firms might be able to produce the industry output vector at the same total cost as a monopolist. But, as Proposition 9 shows, the region in which ray average costs are *strictly* decreasing must only extend by any

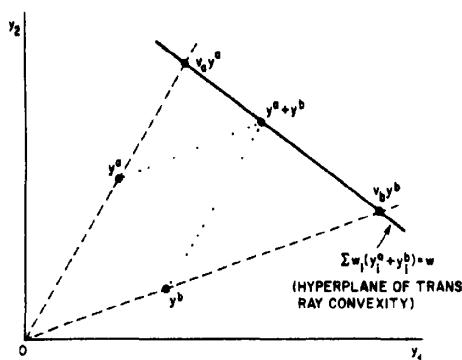



FIGURE 5

small amount beyond the *midpoint* of the ray between the origin and the outermost portion of the ray that is being tested for subadditivity.

The reason we do not need *strict* transray convexity also indicates why the conditions of Proposition 12 are not *necessary* for subadditivity. Let y be the vector sum of two output vectors y^a and y^b. Then y will be further from the origin than either y^a and y^b, and it will lie on a ray between the two. That is, it involves a larger scale than either of the component vectors *as well as* simultaneous production of the different commodities they include. Thus, even if there is no complementarity in the sense of strict transray convexity, if economies of scale are sufficiently strong, production of y by a single firm may still be the least expensive way to carry out the task.[11]

The fact that the conditions of Proposition 12 are not necessary implies that other and, perhaps, more familiar sufficient cost conditions for subadditivity may yet be found. It should be noted, however, that this may not be an important issue for empirical work in which data are likely to be analyzed with the aid of relatively simple mathematical forms for the cost function, whose subadditivity (or its absence) can be judged more directly.

[11]Indeed, it can be shown that mild transray *concavity* can be compatible with subadditivity—the more rapid the rate of decline of ray average costs, the greater the permissible degree of concavity. The proof is a straightforward extension of the proof of Proposition 12. It can also be shown that if some outputs (or every output) in the industry product mix have their own fixed cost levels (a violation of transray convexity at the axes), then the cost function will still be subadditive if the conditions of Proposition 12 hold for the variable cost. The reason is that such fixed costs are always subadditive, and summation of two functions, one subadditive and the other strictly subadditive, always yields a strictly subadditive function.

VIII. Concluding Comment

The implications of the results of the preceding section should not be underestimated. It may seem as though ray subadditivity by itself is enough to constitute evidence for the presence of significant elements of natural monopoly in an industry and that transray subadditivity is little more than a bit of supplementary information. More specifically, it may seem that while transray subadditivity is required for natural monopoly in the production of several commodities together, ray subadditivity by itself gives us the essence of the natural monopoly, for then *each* good is still produced most cheaply by a single firm.

But that conclusion is quite misleading. To illustrate the point, suppose that the industry in question produces two commodities y_1 and y_2, and that, as in the case of Figure 6, a transray cross section has a number of minima. Then, despite strictly declining ray average costs, it can be cheapest for production to be carried out simultaneously by two or more firms. Though each firm will have a "monopoly" of production of the output *proportions* corresponding to its own ray, each will be producing the same commodities as the others. What sorts of monopolies would these really be if one were to "specialize" the production of y_1 and y_2 in a 50–50 ratio and another in a 45–55 ratio? Each firm would be producing the same two goods, only in (slightly?) different proportions, and that is surely no monopoly at all. In other words, *evidence of strict ray subadditivity is, by itself, evidence of virtually nothing of substance so far as natural monopoly is concerned.* It is consistent with the presence of a large number of firms in a cost-minimizing arrangement for the industry, and only requires slight dissimilarities in their product mix. It should

FIGURE 6

be noted that this is no extraordinary or pathological state of affairs. On the contrary, it is frequently found in multiproduct oligopoly industries in reality.[12]

We may note, finally, that though our natural monopoly criteria do not make for easy empirical testing, it is by no means unmanageable. Thus, as part of our research project on the cost of information supply, we have already found it possible with the aid of Proposition 12 to carry out tests of the hypothesis that there is subadditivity in the provision of a number of scientific journals by a single publisher (see the author and Braunstein).

[12]It can be proved that if the cost function is ray concave everywhere but not subadditive, then the number of firms that minimize the cost of producing the industry output vector will never exceed n, the number of different products supplied by the industry. This by itself means that in most multiproduct industries even where costs satisfy the ray concavity requirement (which is stronger than the conventional concept of scale economies) it is perfectly possible to have dozens of firms corresponding to the dozens of products supplied by the industry. Nor need these firms specialize in the production of single products. Rather, it may be optimal for them to operate along rays in the interior of product space, depending on the transray behavior of the cost surface, as is indicated in Figure 6 where it may obviously be more economical for two firms to operate at the low points of the scalloped cross sections than above the axes. Moreover, if, on the usual (but not quite accurate) criterion of scale economies, the cost function has declining ray average costs but is not concave everywhere, then the cost-minimizing number of firms can even exceed the number of products of the industry. Examples showing this, and other proofs underlying this discussion, were provided by Thijs ten Raa and Dietrich Fischer. These materials are presented in the Appendix (Propositions 13 and 14).

APPENDIX

On the Cost-Minimizing Number of Firms When Ray Average Costs Decline

I conclude with two propositions which indicate how much (or how little) we know about the natural number of firms in an industry, given information only about cost behavior along each ray (economies of scale) but no information on transray cost behavior (economies of scope) (see also the author and Fischer).

PROPOSITION 13 (Thijs ten Raa): *If the cost function is strictly output-ray concave, then the optimal number of firms cannot exceed n, the number of commodities produced by the industry.*[13]

PROOF:

It is sufficient to prove that for every $(n+1)$-tuple of output vectors $y^1, \dots, y^{n+1}$ there exists a cheaper n-tuple of output vectors with the same total output. Because in n-dimensional space any $n+1$ vectors are linearly dependent, $y^1, \dots, y^{n+1}$ must be linearly dependent. Hence, one of them, say, without loss generality y^{n+1} is a linear combination of the others:

$$y^{n+1} = \sum_{i=1}^{n} C_i y^i, \quad C_i \geq 0, \quad \text{not all zero}$$

Let $$\lambda \epsilon \left[0, \min_{i=1,\dots,n} \left(1 + \frac{1}{C_i}\right)\right]$$

then

$$\forall i \epsilon \{1, \dots, n\}: (1 + C_i - \lambda C_i) y^i, \quad \lambda y^{n+1} \epsilon \overline{\mathcal{R}_n^+}$$

[13]The proposition holds even if the concavity is *not* strict, provided there is no degeneracy in the sense used in linear programming.

and the sum of these $n + 1$ vectors equals

$$\sum_{i=1}^{n} y^i + \sum_{i=1}^{n} (C_i - \lambda C_i) y^i + \lambda \sum_{i=1}^{n} C_i y^i = \sum_{i=1}^{n} y^i + \sum_{i=1}^{n} C_i y^i = \sum_{i=1}^{n+1} y^i$$

C is output ray concave, hence

$$\forall i \epsilon \{1, \ldots, n\}: \; C\{(1 + C_i - \lambda C_i) y^i\}$$

is a concave function of λ and so is $C(\lambda y^{n+1})$. Hence

$$\sum_{i=1}^{n} C\{(1 + C_i - \lambda C_i) y^i\} + C(\lambda y^{n+1})$$

is a concave function of λ. Hence its minimum occurs at one of the end points of λ's range, that is, at

$$\lambda = 0 \quad \text{or } \lambda = \min_{i=1,\ldots,n} \left(1 + \frac{1}{C_i}\right) = 1 + \frac{1}{C_j} \quad \text{for some } j \epsilon \{1, \ldots, n\}$$

When $\lambda = 0$, then the $(n + 1)$th firm does not produce anything. When $\lambda = 1 + 1/C_j$, then the j th firm does not produce anything. In either case only n firms are left, producing the industry output more cheaply than when $\lambda = 1$ under which we have the original outputs $y^1, \ldots, y^{n+1}$. However,

PROPOSITION 14 (Raa and Fischer): *Declining ray average costs alone do not preclude the optimality of a number of firms larger than the number of products supplied by the industry.*

PROOF:

The following is an example in which it is optimal to have three firms producing two commodities when ray average costs are (not strictly) declining and there are no fixed costs.[14] Let

$$y^I = \left(1 + \frac{\sqrt{3}}{3}, 2\right)$$

and, to simplify the argument (letting us deal with only three rays in output space) let C be such that production is relatively cheap along rays involving three particular output bundles

$$(y_2 \equiv 0, y_1 \equiv 0 \quad \text{and} \quad y_2 = \sqrt{3} y_1)$$

but prohibitively expensive along all other rays. In addition, let the cost function satisfy

$$C\left(\frac{y_2\sqrt{3}}{3}, y_2\right) = 2y_2 \quad \text{for} \quad y_2 \geq 0$$

$$C(0, y_2) = \sqrt{3}\, y_2 \quad \text{for} \quad y_3 \geq 0$$

$$C(y_1, 0) = y_1 \quad \text{for} \quad y_1 \geq 1; \sqrt{3} - 1 < C\left(1 - \frac{\sqrt{3}}{3}, 0\right) < 1$$

$$C(y_1, 0) \quad \text{is linear for} \quad y_1 \epsilon \left[0, 1 - \frac{\sqrt{3}}{3}\right]$$

$$\text{and for} \quad y_1 \epsilon \left[1 - \frac{\sqrt{3}}{3}, 1\right]$$

Then y^I is produced more cheaply by

$$y^1 = (1, 0), y^2 = \left(\frac{\sqrt{3}}{3}, 1\right) \text{ and } y^3 = (0, 1)$$

To show this, because of the linearity of C, it is sufficient to show that (y^1, y^2, y^3) is cheaper than

$$\left(\left(1 - \frac{\sqrt{3}}{3}, 0\right), \left(\frac{2\sqrt{3}}{3}, 2\right)\right)$$

and

$$\left(\left(1 + \frac{\sqrt{3}}{3}, 0\right), (0, 2)\right)$$

the two pairs of output vectors capable of producing y^I. Thus:

$$C(y^1) + C(y^2) + C(y^3) = 1 + 2 + \sqrt{3} = 3 + \sqrt{3}$$

$$C\left(1 - \frac{\sqrt{3}}{3}, 0\right) + C\left(\frac{2\sqrt{3}}{3}, 2\right) > \sqrt{3} - 1 + 4 = 3 + \sqrt{3}$$

$$C\left(1 + \frac{\sqrt{3}}{3}, 0\right) + C(0, 2) = 1 + \frac{\sqrt{3}}{3} + 2\sqrt{3} = 1 + \left(2\frac{1}{3}\right)\sqrt{3} > 3 + \sqrt{3}$$

[14]Note that the addition of any positive fixed cost will make the ray average costs decline *strictly*, without affecting the example.

REFERENCES

W. J. Baumol, "Scale Economies, Average Cost and the Profitability of Marginal-Cost Pricing," in Ronald E. Grieson, ed., *Essays in Urban Economics and Public Finance in Honor of William S. Vickrey*, Lexington 1976.

———, **E. E. Bailey, and R. D. Willig**, "Weak Invisible Hand Theorems on Pricing and Entry in a Multiproduct Natural Monopoly," *Amer. Econ. Rev.*, June 1977, *67*, 350–65.

——— **and D. Fischer**, "On the Optimal Number of Firms in an Industry," *Quart. J. Econ.*, 1977, forthcoming.

——— **and Y. M. Braunstein**, "Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication," *J. Polit. Econ.*, 1977, forthcoming.

G. R. Faulhaber, (1975a) "Cross-Subsidization: Pricing in Public Enterprise," *Amer. Econ. Rev.*, Dec. 1975, *65*, 966–77.

———, (1975b) "Pricing Rules in Cooperative Markets," unpublished doctoral dissertation, Princeton Univ. 1975.

——— **and E. E. Zajac**, "Some Thoughts on Cross-Subsidization in Regulated Industries," paper presented at the International Conference in Telecommunications, England 1974.

G. Hanoch, "Homotheticity in Joint Production," *J. Econ. Theory*, Dec. 1970, *2*, 423–26.

———, "The Elasticity of Scale and the Shape of Average Costs," *Amer. Econ. Rev.*, June 1975, *65*, 492–97.

E. D. Lowry, "Justification for Regulation: The Case for Natural Monopoly," *Publ. Util. Fortnightly*, Nov. 8, 1973, *28*, 1–7.

P. Newman, "Some Properties of Concave Functions," *J. Econ. Theory*, Oct. 1969, *1*, 291–314.

J. C. Panzar and R. D. Willig, (1975a) "Economies of Scale and Economies of Scope in Multi-Output Production," unpublished paper, Bell Laboratories 1975.

——— **and** ———, (1975b) "Free Entry and the Sustainability of Natural Monopoly," unpublished paper, Bell Laboratories 1975.

R. A. Posner, "Natural Monopoly and its Regulation," *Stanford Law Rev.*, Feb. 1969, *21*, 548–643.

R. S. Seneca, "Inherent Advantage, Costs and Resource Allocation in the Transportation Industry," *Amer. Econ. Rev.*, Dec. 1973, *63*, 945–56.

W. W. Sharkey and L. G. Telser, "Supportable Cost Functions for the Multiproduct Firm," unpublished paper, Bell Laboratories 1977.

Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods

By R. DORNBUSCH, S. FISCHER, AND P. A. SAMUELSON*

This paper discusses Ricardian trade and payments theory in the case of a continuum of goods. The analysis thus extends the development of many-commodity, two-country comparative advantage analysis as presented, for example, in Gottfried Haberler (1937), Frank Graham (1923), Paul Samuelson (1964), and Frank W. Taussig (1927). The literature is historically reviewed by John Chipman (1965). Perhaps surprisingly, the continuum assumption simplifies the analysis neatly in comparison with the discrete many-commodity case. The distinguishing feature of the Ricardian approach emphasized in this paper is the determination of the competitive margin in production between imported and exported goods. The analysis advances the existing literature by formally showing precisely how tariffs and transport costs establish a range of commodities that are not traded, and how the price-specie flow mechanism does or does not give rise to movements in relative cost and price levels.

The formal *real* model is introduced in Section I. Its equilibrium determines the *relative* wage and price structure and the efficient international specialization pattern. Section II considers standard comparative static questions of growth, demand shifts, technological change, and transfers. Extensions of the model to nontraded goods, tariffs, and transport costs are then studied in Section III. Monetary considerations are introduced in Section IV, which examines the price-specie mechanism under stable parities, floating exchange rate regimes, and also questions of unemployment under sticky money wages.

*Massachusetts Institute of Technology. Helpful comments from Ronald W. Jones are gratefully acknowledged. Financial support was provided by a Ford Foundation grant to Dornbusch, NSF GS-41428 to Fischer, and NSF 75-04053 to Samuelson.

I. The Real Model

In this section we develop the basic real model and determine the equilibrium relative wage and price structure along with the efficient geographic pattern of specialization. Assumptions about technology are specified in Section IA. Section IB deals with demand. In Section IC the equilibrium is constructed and some of its properties are explored. Throughout this section we assume zero transport costs and no other impediments to trade.

A. *Technology and Efficient Geographic Specialization*

The many-commodity Ricardian model assumes constant unit labor requirements $(a_1, \ldots, a_n)$ and $(a_1^*, \ldots, a_n^*)$ for the n commodities that can be produced in the home and foreign countries, respectively. The commodities are conveniently indexed so that relative unit labor requirements are ranked in order of diminishing home country comparative advantage,

$$a_1^*/a_1 > \ldots > \ldots > a_i^*/a_i > \ldots > a_n^*/a_n$$

where an asterisk denotes the foreign country.

In working with a continuum of goods, we similarly index commodities on an interval, say $[0, 1]$, in accordance with diminishing home country comparative advantage. A commodity z is associated with each point on the interval, and for each commodity there are unit labor requirements in the two countries, $a(z)$ and $a^*(z)$, with relative unit labor requirement given by

$$(1) \qquad A(z) \equiv \frac{a^*(z)}{a(z)} \qquad A'(z) < 0$$

The relative unit labor requirement function in (1) is by strong assumption continuous,

and by construction (ranking or indexing of goods), decreasing in z. The function $A(z)$ is shown in Figure 1 as the downward sloping schedule.

Consider now the range of commodities produced domestically and those produced abroad, as well as the relative price structure associated with given wages. For that purpose we define as w and w^* the domestic and foreign wages measured in *any* (common!) unit. The home country will efficiently produce all those commodities for which domestic unit labor costs are less than or equal to foreign unit labor costs. Accordingly, any commodity z will be produced at home if

$$a(z)w \leq a^*(z)w^* \tag{2}$$

Thus

$$\omega \leq A(z) \tag{2'}$$

where (3) defines the parameter ω, fundamental to Ricardian analysis,

$$\omega \equiv w/w^* \tag{3}$$

This is the ratio of our real wage to theirs (our "double-factoral terms of trade"). It follows that for a given relative wage ω the home country will efficiently produce the range of commodities

$$0 \leq z \leq \tilde{z}(\omega) \tag{4}$$

where taking (2′) with equality defines the borderline commodity z, for which

$$\tilde{z} = A^{-1}(\omega) \tag{5}$$

$A^{-1}(\ )$ being the inverse function of $A(\ )$. By the same argument the foreign country will specialize in the production of commodities in the range

$$\tilde{z}(\omega) \leq z \leq 1 \tag{4'}$$

The minimum cost condition determines the structure of relative prices. The relative price of a commodity z in terms of any other commodity z', when both goods are produced in the home country, is equal to the ratio of home unit labor costs:

$$P(z)/P(z') = wa(z)/wa(z') = a(z)/a(z'); \quad z \leq \tilde{z}, z' \leq \tilde{z} \tag{6}$$

The relative price of home produced z in terms of a commodity z'' produced abroad is by contrast

$$P(z)/P(z'') = wa(z)/w^*a^*(z'') = \omega a(z)/a^*(z''); \quad z < \tilde{z} < z'' \tag{7}$$

In summarizing the supply part of the model we note that any specified relative real wage is associated with an efficient geographic specialization pattern characterized by the borderline commodity $\tilde{z}(\omega)$ as well as by a relative price structure. (The pattern is "efficient" in the sense that the world is out on, and not inside, its production-possibility frontier.)

B. *Demand*

On the demand side, the simplest Mill-Ricardo analysis imposes a strong homothetic structure in the form of J. S. Mill or Cobb-Douglas demand functions that associate with each ith commodity a *constant expenditure share*, b_i. It further assumes *identical* tastes for the two countries or *uniform* homothetic demand.

By analogy with the many-commodity case, which involves budget shares

$$b_i = P_iC_i/Y \qquad b_i = b_i^*$$

$$\sum_1^n b_i = 1$$

We therefore prescribe for the continuum case a given $b(z)$ profile:

$$b(z) = P(z)C(z)/Y > 0 \tag{8}$$
$$b(z) = b^*(z)$$
$$\int_0^1 b(z)\,dz = 1$$

where Y denotes total income, C demand for and P the price of commodity z.

Next we define the fraction of income spent (anywhere) on those goods in which the home country has a comparative advantage:

$$\vartheta(\tilde{z}) \equiv \int_0^{\tilde{z}} b(z)\,dz > 0 \tag{9}$$
$$\vartheta'(\tilde{z}) = b(\tilde{z}) > 0$$

where again $(0, \bar{z})$ denotes the range of commodities for which the home country enjoys a comparative advantage. With a fraction ϑ of each country's income, and therefore of world income, spent on home produced goods, it follows that the fraction of income spent on foreign produced commodities is

$$(9') \quad 1 - \vartheta(\bar{z}) \equiv \int_{\bar{z}}^{1} b(z)dz$$

$$0 \leq \vartheta(z) \leq 1$$

C. *Equilibrium Relative Wages and Specialization*

To derive the equilibrium relative wage and price structure and the associated pattern of efficient geographic specialization, we turn next to the condition of market equilibrium. Consider the home country's labor market, or equivalently the market for domestically produced commodities. With $\bar{z}$ denoting the *hypothetical* dividing line between domestically and foreign produced commodities, equilibrium in the market for home produced goods requires that domestic labor income wL equals world spending on domestically produced goods:

$$(10) \qquad wL = \vartheta(\bar{z})(wL + w^*L^*)$$

Equation (10) associates with each $\bar{z}$ a value of the relative wage w/w^* such that market equilibrium obtains. This schedule is drawn in Figure 1 as the upward sloping locus and is obtained from (10) by rewriting the equation in the form:

$$(10') \quad \omega = \frac{\vartheta(\bar{z})}{1 - \vartheta(\bar{z})}(L^*/L) = B(\bar{z}; L^*/L)$$

where it is apparent from (9) that the schedule starts at zero and approaches infinity as $\bar{z}$ approaches unity.

To interpret the $B(\ )$ schedule we note that it is entirely a representation of the demand side; and in that respect it shows that if the range of domestically produced goods were increased at constant relative wages, demand for domestic labor (goods) would increase as the dividing line is shifted —at the same time that demand for foreign

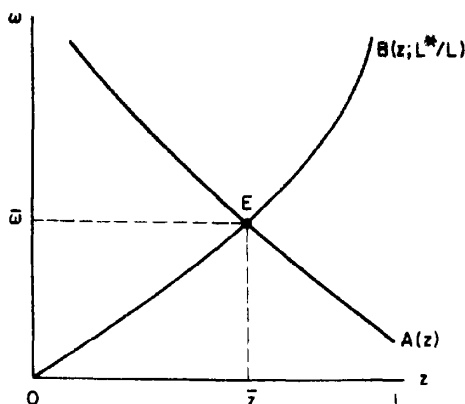



FIGURE 1

labor (goods) would decline.[1] A rise in the domestic relative wage would then be required to equate the demand for domestic labor to the existing supply.

An alternative interpretation of the $B(\ )$ schedule as the locus of trade balance equilibria uses the fact that (10) can be written in the balance-of-trade form:

$$(10'') \quad [1 - \vartheta(\bar{z})]wL = \vartheta(\bar{z})w^*L^*$$

This states that equilibrium in the trade balance means imports are equal in value to exports. On this interpretation, the $B(\ )$ schedule is upward sloping because an increase in the range of commodities hypothetically produced at home at constant relative wages lowers our imports and raises our exports. The resulting trade imbalance would have to be corrected by an increase in our relative wage that would raise our import demand for goods and reduce our exports, and thus restore balance.

The next step is to combine the demand side of the economy with the condition of efficient specialization as represented in equation (5), which specifies the competitive margin as a function of the relative wage. Substituting (5) in (10') yields as a solution the unique relative wage $\bar{\omega}$, at which the world is efficiently specialized, is in bal-

[1]Throughout this paper we refer to "domestic" goods as commodities produced in the home country rather than to commodities that are nontraded. The latter we call "nontraded" goods.

anced trade, and is at full employment with all markets clearing:

$$\bar{\omega} = A(\bar{z}) = B(\bar{z}; L^*/L) \tag{11}$$

The equilibrium relative wage defined in (11) is represented in Figure 1 at the intersection of the $A(\ )$ and $B(\ )$ schedules.[2] Commodity $\bar{z}$ denotes the equilibrium borderline of comparative advantage between commodities produced and exported by the home country ($0 \leq z \leq \bar{z}$), and those commodities produced and exported by the foreign country ($\bar{z} \leq z \leq 1$).

Among the characteristics of the equilibrium we note that the equilibrium relative wages and specialization pattern are determined by technology, tastes, and relative size (as measured by the relative labor force).[3] The relative price structure associated with the equilibrium at point E is defined by equations (6) and (7) once (11) has defined the relative wage $\bar{\omega}$ and the equilibrium specialization pattern $\bar{z}(\bar{\omega})$.

The equilibrium levels of production $Q(z)$ and $Q^*(z)$, and employment in each industry $L(z)$ and $L^*(z)$, can be recovered from the demand structure and unit labor requirements once the comparative advantage pattern has been determined.

We note that with identical homothetic tastes across countries and no distortions, the relative wage $\bar{\omega}$ is a measure of the well-being of the representative person-laborer at home relative to the well-being of the representative foreign laborer.

II. Comparative Statics

The unique real equilibrium in Figure 1 is determined jointly by tastes, technology, and relative size, L^*/L. We can now exploit Figure 1 to examine simple comparative static questions.

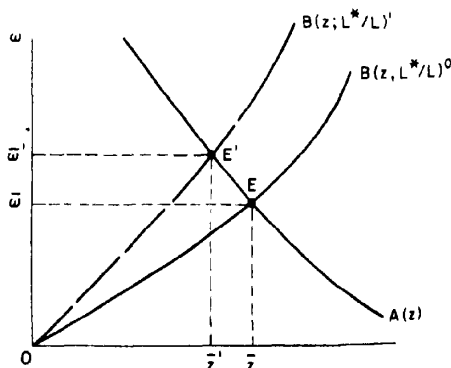
A. *Relative Size*

Consider first the effect of an increase in the relative size of the rest of the world. An increase in L^*/L by (10) shifts the $B(\ )$ trade balance equilibrium schedule upward in proportion to the change in relative size and must, therefore, raise the equilibrium relative wage at home and reduce the range of commodities produced domestically. It is apparent from Figure 2 that the domestic relative wage increases *proportionally less* than the decline in domestic relative size.

The rise in equilibrium relative wages due to a change in relative size can be thought of in the following manner. At the initial equilibrium, the increase in the foreign relative labor force would create an excess supply of labor abroad and an excess demand for labor at home—or, correspondingly, a trade surplus for the home country. The resulting increase in domestic relative wages serves to eliminate the trade surplus while

FIGURE 2

[2]See the Appendix for the relation of the diagram to previous analyses

[3]The construction of the $B(\ )$ schedule relies heavily on the Cobb-Douglas demand structure. If, instead, demand functions were identical across countries and homothetic, an analogous schedule could be constructed. In the general homothetic case, however, a set of relative prices is required at each z to calculate the equivalent of the $B(\ )$ schedule; the relative prices are those that apply on the $A(\bar{z})$ schedule for that value of z. In this case the independence of the $A(\ )$ and $B(\ )$ schedules is obviously lost. In the general homothetic case there is still a unique intersection of the $A(\ )$ and $B(\ )$ schedules. For more general nonhomothetic demand structures, it is known that an equilibrium exists; but even in the case of two Ricardian goods there may be *no* unique equilibrium even though there will almost always be a finite number of equilibria. See Gerard Debreu and Stephen Smale. Extensions of our analysis with respect to the demand structure and the number of countries are developed in unpublished work by Charles Wilson.

at the same time raising relative unit labor costs at home. The increase in domestic relative unit labor costs in turn implies a loss of comparative advantage in marginal industries and thus a needed reduction in the range of commodities produced domestically.

The welfare implications of the change in relative size take the form of an unambiguous improvement in the home country's real income and (under Cobb-Douglas demand) a reduction in real income per head abroad. We observe, too, that from the definition of the home country's share in world income and (10), we have

$$wL/(wL + w^*L^*) = \vartheta(\bar{z}) \tag{12}$$

It is apparent, as noted above, that a reduction in domestic relative size in raising the domestic relative wage (thereby reducing the range of commodities produced domestically) must under our Cobb-Douglas demand assumptions lower the home country's share in total world income and spending—even though our per capita income rises.

B. *Technical Progress*

To begin with, we are concerned with the effects of uniform technical progress. By equation (1), a uniform proportional reduction in foreign unit labor requirements implies a reduction in $a^*(z)$ and therefore a proportional downward shift of the $A(z)$ schedule in Figure 1. At the initial relative wage $\bar{\omega}$, the loss of our comparative advantage due to a reduction in foreign unit labor costs will imply a loss of some industries in the home country and a corresponding trade deficit. The resulting induced decline in the equilibrium relative wage serves to restore trade balance equilibrium, and to offset in part our decline in comparative advantage.

The net effect is therefore a reduction in domestic relative wages, which must fall proportionally short of the decline in relative unit labor requirements abroad. The home country's terms of trade therefore improve as can be noted by using (7) for any two commodities z and z'', respectively, produced at home and abroad:

$$\hat{P}(z) - \hat{P}(z'') = \hat{\bar{\omega}} - \hat{a}^*(z'') > 0 \tag{13}$$

where a "hat" denotes a proportional change. Domestic real income increases, as does foreign real income.[4] The range of goods produced domestically declines since domestic labor, in efficiency units, is now relatively more scarce.

An alternative form of technical progress that can be studied is the international transfer of the least cost technology. Such transfers reduce the discrepancies in relative unit labor requirements—by lowering them for each z in the relatively less efficient country—and therefore flatten the $A(z)$ schedule in Figure 1. It can be shown that such harmonization of technology must benefit the innovating low-wage country, and that it may reduce real income in the high-wage country whose technology comes to be adopted. In fact, the high-wage country must lose if harmonization is complete so that relative unit labor requirements now become identical across countries and all our consumer's surplus from international trade vanishes.[5]

C. *Demand Shifts*

The case with a continuum of commodities requires a careful definition of a demand shift. For our purposes it is sufficient to ask: What is the effect of a shift from high z commodities toward low z commodities? It is apparent from Figure 2 that such a shift will cause the trade balance equilibrium schedule $B(\ )$ to shift up and to the left. It follows that the equilibrium domestic rela-

[4]The purchasing power of foreign labor income in terms of domestically produced goods is $w^*L^*/wa(z) = L^*/a(z)\bar{\omega}$ and in terms of foreign goods $L^*/a^*(z)$. The fact that foreigners' real income per head rises is guaranteed by our Cobb-Douglas demand assumption. In the general homothetic case, a balanced reduction in $a^*(z)$ can be immiserizing abroad if the real wage falls strongly in terms of all previously imported goods; however, the balanced drop in $a^*(z)$ in the general homothetic case *always* increases our real wage.

[5]Complete equilization of unit labor requirements implies that the $A(\ )$ schedule is horizontal at the level $\omega = A(z) = 1$. In this case geographic specialization becomes indeterminate and inessential.

tive wage will rise while the range of commodities produced by the home country declines. Domestic labor is allocated to a narrower range of commodities that are consumed with higher density while foreign labor is spread more thinly across a larger range of goods.

Welfare changes cannot be identified in this instance because tastes themselves have changed. It is true that domestic relative income rises along with the relative wage. Further we note that since $\bar{\omega}$ rises, the relative well-being of home labor to foreign labor (reckoned at the new tastes) is greater than was our laborers' relative well-being (reckoned at the old tastes).

D. *Unilateral Transfers*

Suppose foreigners make a continual unilateral transfer to us. With uniform homothetic tastes and no impediments to trade, neither curve is shifted by the transfer since we spend the transfer *exactly* as foreigners would have spent it but for the transfer. The new equilibrium involves a recurring trade deficit for us, equal to the transfer, but there is no change in the terms of trade. As Bertil Ohlin argued against John Maynard Keynes, here is a case where full equilibration takes place solely as a result of the spending transfers. When we introduce nontraded goods below, Ohlin's presumption will be found to require detailed qualifications, as it also would if tastes differed geographically.

III. Extensions of the Real Model

Extensions of the real model taken up in this section concern nontraded goods, tariffs, and transport costs. The purpose of this section is twofold. First we establish how the exogenous introduction of nontraded goods qualifies the preceding analysis. Next we turn to a particular specification of tariffs and transport costs to establish an equilibrium range of endogenously determined nontraded goods as part of the equilibrium solution of the model. Transfers are then shown to affect the equilibrium relative price structure and the range of goods traded.

A. *Nontraded Goods*

To introduce nontraded goods into the analysis we assume that a fraction k of income is everywhere spent on internationally traded goods, and a fraction $(1 - k)$ is spent in each country on nontraded commodities. With $b(z)$ continuing to denote expenditure densities for traded goods, we have accordingly

$$k \equiv \int_0^1 b(z)\,dz < 1 \tag{14}$$

where z denotes traded goods.[6] As before the fraction of income spent on domestically exportable commodities is $\vartheta(z)$, except that ϑ now reaches a maximum value of $\vartheta(1) = k$.

Equation (1) remains valid for traded goods, but the trade balance equilibrium condition in (10″) must now be modified to:

$$[1 - \vartheta(\bar{z}) - (1 - k)]wL = \vartheta(\bar{z})w^*L^* \tag{15}$$

since domestic spending on imports is equal to income less spending on *all* domestically produced goods including nontraded commodities. Equation (15) can be rewritten as

$$\omega = \frac{\vartheta(\bar{z})}{k - \vartheta(\bar{z})}(L^*/L) \tag{15'}$$

where k is a constant and therefore independent of the relative wage structure.

We note that (15′) together with (5) determines the equilibrium relative wage and efficient geographic specialization, $(\bar{\omega}, \bar{z})$. Further it is apparent that (15′) has exactly the same properties as (10′) and that accordingly a construction of equilibrium like that in Figure 1 remains appropriate. The equilibrium relative wage again depends on

[6]We can think of the range of nontraded goods as another [0, 1] interval with commodities denoted by x and expenditure fractions on those goods given by $c(x)$. With these definitions we have $\int_0^1 c(x)\,dx = 1 - k$, a positive fraction.

relative size, technology, and demand conditions. In this case demand conditions explicitly include the fraction of income spent on traded goods:

$$(11') \qquad \bar{\omega} = \frac{\vartheta(\bar{z})}{k - \vartheta(\bar{z})} \frac{L^*}{L} = A(\bar{z})$$

This nicely generalizes our previous equilibrium of (11) to handle exogenously given nontraded goods.[7]

Two applications of the extended model highlight the special aspects newly introduced by nontraded goods. First consider a shift in demand (in each country) toward *nontraded* goods. To determine the effects on the equilibrium relative wage we have to establish whether this shift is at the expense of high or low z commodities. In the former case the home country's relative wage increases while in the latter case it declines. If the shift in demand in each country is uniform so that $b(z)$ is reduced in the same proportion for all z in both countries, then the relative wage remains unchanged.

Consider next a transfer received by the home country in the amount T measured in terms of foreign labor. As is well known, and already shown, with identical homothetic tastes and *no* nontraded goods, a transfer leaves the terms of trade unaffected. In the present case, however, the condition for balanced trade, inclusive of transfers, becomes:

$$(16) \qquad T = (k - \vartheta)[\omega L + T] - \vartheta[L^* - T]$$

or, in equilibrium,

$$(16') \qquad \bar{\omega} = \frac{1 - k}{k - \vartheta(\bar{z})}(T/L) + \frac{\vartheta(\bar{z})}{k - \vartheta(\bar{z})}(L^*/L)$$

It is apparent from (16') that a transfer receipt by the home country causes the trade balance equilibrium schedule in Figure 1 to shift upward at each level of z. Accordingly, the equilibrium domestic relative wage increases and the range of commodities produced domestically is reduced. The steps in achieving this result are, first, that at the initial relative wage only a fraction of the transfer is spent on imports in the home country, while foreign demand for domestic goods similarly declines only by a fraction of their reduced income. The resulting surplus for the home country has to be eliminated by, second, an increase in the domestic relative wage and a corresponding improvement in the home country's terms of trade.[8]

The analysis of nontraded goods therefore confirms in a Ricardian model the "orthodox" presumption with respect to the terms of trade effects of transfers.[9]

B. *Transport Costs: Endogenous Equilibrium for Nontraded Goods*

The notion that transport costs give rise to a range of commodities that are nontraded is established in the literature and is particularly well stated by Haberler (1937). In contrast with the previous section we shall now endogenously determine the range of nontraded commodities as part of the equilibrium. We assume, following the "iceberg" model of Samuelson (1954), that transport costs take the form of "shrinkage" in transit so that a fraction $g(z)$ of commodity z shipped actually arrives. We further impose the assumption that $g = g(z)$ is identical for all commodities and the same for shipments in either direction.

The home country will produce commodities for which domestic unit labor cost falls short of foreign unit labor costs adjusted for shrinkage, and we modify (2') accordingly:

$$(17) \qquad wa(z) \leq (1/g)w^*a^*(z)$$

or

$$\omega \leq A(z)/g$$

[7]Diagrams much like Figures 1 and 2 again apply: the descending $A(z)$ schedule is as before; and now the new rising schedule looks much as before. As before, a rise in L^*/L and a balanced drop in $a^*(z)$ will raise $\bar{\omega}$ and lower $\bar{z}$.

[8]At constant relative wages the current account worsens by $[(1 - k - \vartheta) + \vartheta]dT = (1 - k)dT$ which is less than the transfer, since it is equal to the fraction of income spent on nontraded goods.

[9]The pre-Ohlin orthodox view of Keynes, Taussig, Jacob Viner and other writers is discussed in Viner (1937) and Samuelson (1952, 1954). A recent treatment with nontraded goods is Ronald Jones (1975).

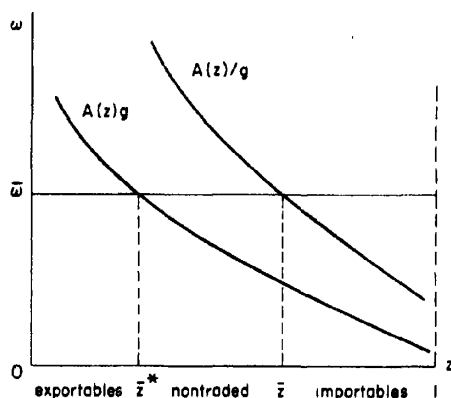



FIGURE 3

Similarly the foreign country produces commodities for which foreign unit labor cost falls short of adjusted unit labor costs of delivered imports:

$$(18) \qquad w^*a^*(z) \leq (1/g)wa(z)$$

$$\text{or} \qquad A(z)g \leq \omega$$

In Figure 3 we show the adjusted relative unit labor requirement schedules $A(z)/g$ and $A(z)g$. It is apparent from (17) and (18) that for any given relative wage the home country produces and exports commodities to the left of the $A(z)g$ schedule, both countries produce as nontraded goods commodities in the intermediate range, and the foreign country produces and exports commodities in the range to the right of $A(z)/g$.

To determine the equilibrium relative wage we turn to the trade balance equilibrium condition in (19)—together with (20) and (21)—which is modified to take account of the endogenous range of nontraded goods:

$$(19) \quad (1-\lambda)wL = (1-\lambda^*)w^*L^*$$

The variable λ is the fraction of home country income spent on our domestically (or home) produced goods—exportables and nontraded—and λ^* is the share of foreigners' income spent on goods they produce. Both λ and λ^* are endogenously determined because the range of goods produced in each country depends on the relative wages.

$$(20) \quad \lambda(g\omega) \equiv \int_0^{\bar{z}} b(z)dz \qquad \lambda'(g\omega) < 0$$

$$\lambda^*(\omega/g) \equiv \int_{\bar{z}}^1 b(z)dz \quad \lambda^{*\prime}(\omega/g) > 0$$

The dependence of $\lambda(\ )$ and $\lambda^*(\ )$ on the variables specified in (20) and the respective derivatives follow from (21) below.

The limits of integration $\bar{z}$ and $\bar{z}^*$ are derived from the conditions for efficient production in (17) and (18) by imposing equalities and so defining the borderline commodities. Thus, in Figure 3, $\bar{z}$ is the borderline between domestic nontraded goods and imports for the home country, and $\bar{z}^*$ denotes the borderline between foreign nontraded goods and the home country's exports:

$$(21) \quad \bar{z}^* = A^{-1}(\omega/g) \quad d\bar{z}^*/d(\omega/g) < 0$$

$$\bar{z} = A^{-1}(g\omega) \qquad d\bar{z}/d(g\omega) < 0$$

Of course, equilibrium $\bar{z}$ and $\bar{z}^*$ are yet to be determined by the interaction of technology and demand conditions.

From (21) an increase in the relative wage reduces the range of commodities domestically produced and therefore raises the fraction of income spent on imports. Abroad the converse holds. An increase in the domestic relative wage increases the range of goods produced abroad and therefore reduces the fraction of income spent on imports. It follows that we can solve:

$$(19') \quad \bar{\omega} = \frac{1-\lambda^*(\bar{\omega}/g)}{1-\lambda(g\bar{\omega})}(L^*/L)$$

$$\equiv \varphi(\bar{\omega}; L^*/L, g) \quad \partial\varphi/\partial\bar{\omega} < 0$$

for the unique equilibrium relative wage as a function of relative size and transport costs:

$$(22) \quad \bar{\omega} = \bar{\omega}(L^*/L, g)$$

Because (19')'s right-hand side declines as $\bar{\omega}$ rises, a rise in L^*/L must still raise $\bar{\omega}$; a rise in g can shift $\bar{\omega}$ in either direction, depending on the $B(z)$ and $A(z)$ profiles.

The equilibrium relative wage in (22), taken in conjunction with (21), determines the equilibrium geographic production pattern, $\bar{z}$ and $\bar{z}^*$. Since the range of nontraded

goods $\bar{z}^* \leq z \leq \bar{z}$ depends in this formulation on the equilibrium relative wage, it is obvious that shifts in given parameters will shift the range of nontraded commodities. Thus, a transfer that raises the equilibrium relative wage at home causes previously exported commodities to become nontraded, and previously nontraded commodities to become importables.

C. *Tariffs*

We consider next the case of zero transport cost but where each country levies a uniform tariff on imports at respective rates t and t^*, with proceeds rebated in lump sum form. This case, too, leads to cost barriers to importing, and to a range of commodities that are not traded, with the boundaries defined by:

$$\bar{z} = A^{-1}\left(\frac{\omega}{1+t}\right) \tag{23}$$

and $\bar{z}^* = A^{-1}(\omega(1+t^*))$

From (23) it is apparent that the presence of tariffs in either or both countries must give rise to nontraded goods because in this case $\bar{z} \neq \bar{z}^*$.

The trade balance equilibrium condition at international prices becomes, in place of (19),

$$(1-\lambda)Y/(1+t) = (1-\lambda^*)Y^*/(1+t^*) \tag{24}$$

where Y and Y^* denote incomes inclusive of lump sum tariff rebates. Using the fact that rebates are equal to the tariff rate times the fraction of income spent on imports, we arrive at the trade balance equilibrium condition in the form:[10]

$$\omega = \left(\frac{1-\lambda^*}{1-\lambda}\right)\frac{1+t\lambda}{1+t^*\lambda^*}(L^*/L) \tag{25}$$

where λ and λ^* are functions of (ω, t, t^*).

[10]Tariff rebates in the home country are equal to $R = (1-\lambda)Yt/(1+t)$. With $Y \equiv WL + R$ we therefore have $Y = WL(1+t)/(1+\lambda t)$ as an expression for income inclusive of transfers. From equations (20) and (23) we have $\lambda = \lambda[\omega/(1+t)]$ and $\lambda^* = \lambda^*[\omega(1+t^*)]$, having substituted the tariff instead of transport costs as the obstacle to trade.

The implicit relations (25) can be solved for the equilibrium relative wage as a function of relative size and the tariff structure:

$$\bar{\omega} = \bar{\omega}(L^*/L, t, t^*) \tag{26}$$

From (26) and (23) it is apparent now that the range of nontraded goods will be a function of both tariff rates. It is readily shown that an increase in the tariff improves the imposing country's relative wage and terms of trade. Furthermore, as is well known, when all countries but one are free traders, then one country can always improve its own welfare by imposing a tariff that is not too large.

A further question suggested by (26) concerns the effect of a uniform increase in world tariffs. Starting from zero, a small uniform increase in tariffs raises the relative wage of the country whose commodities command the larger share in world spending. This result occurs for two reasons. First, at the initial relative wage a larger share of spending out of tariff rebates falls on the goods of the country commanding a larger share in world demand. Second, the tariff induces new nontraded goods and therefore increases net demand for the borderline commodity of the country whose residents have the larger income, or equivalently, the larger share in world income.

If countries are of equal size as measured by the share in world income, such a uniform tariff increase has zero effect on relative wages, but of course reduces well-being in both places. Multilateral tariff increases, in this case, unnecessarily create some nontraded goods, and artificially raise the relative price of importables in terms of domestically produced commodities in each country exactly in proportion to the tariff.

IV. Money, Wages, and Exchange Rates

In this section we extend the discussion of the Ricardian model to deal with monetary aspects of trade. Specifically we shall be interested in the determination of exchange rates in a flexible rate system, in the process of adjustment to trade imbalance under fixed rates, and in the role of wage sticki-

ness. The purpose of the extension is to integrate real and monetary aspects of trade.

A. *Flexible Exchange Rates*

The barter analysis of the preceding sections is readily extended to a world of flexible exchange rates and flexible money wages. Assume a given nominal quantity of money in each country, M and M^*, respectively. Further, in accordance with the classical Quantity Theory, assume constant expenditure velocities V and V^*.[11] A flexible exchange rate, and our stipulating the absence of nonmonetary international asset flows, will assure trade balance equilibrium and therefore the equality of income and spending in each country. The nominal money supplies and velocities determine nominal income in each country:

$$(27) \quad WL = MV \quad \text{and} \quad W^*L^* = M^*V^*$$

where W and W^* (now in capital letters) denote domestic and foreign money wages in terms of the respective currencies. Further, defining the exchange rate e as the domestic currency price of foreign exchange, the foreign wage measured in terms of domestic currency is eW^*, and the relative wage therefore is $\omega \equiv W/eW^*$.

From the determination of the equilibrium real wage ratio $\bar{\omega}$ by our earlier "real" relations, we can now find an expression for the equilibrium exchange rates:

$$(28) \quad \bar{e} = (1/\bar{\omega})(\overline{W}/\overline{W}^*) = (1/\bar{\omega})(MV/M^*V^*)(L^*/L)$$

where (27′) defines equilibrium money wages:

$$(27') \qquad \overline{W} = MV/L$$

and

$$\overline{W}^* = M^*V^*/L^*$$

In this simple structure and with wage flexibility, we can keep separate the determinants of all equilibrium real variables from all monetary considerations. Monetary changes or velocity changes in one country will be reflected in equiproportionate changes in prices in that country and in the exchange rate in the fashion of the neutral-money Quantity Theory. However, a real disturbance, as (28) shows, definitely does have repercussions on the nominal exchange rate as well as on the real equilibrium.

Using the results of Section II, we see that an increase in the foreign relative labor force causes, under flexible exchange rates and given $\overline{M}$ and $\overline{M}^*$, a depreciation in the home country's exchange rate as does uniform technical progress abroad. A shift in real demand toward foreign goods likewise leads to a depreciation of the exchange rate as well as to a reduction in real $\bar{\omega}$. A rise in foreign tariffs will also cause our currency to depreciate. Each of these real shifts is assumed to take place while (M, M^*) are unchanged and on the simplifying proviso that real income changes leave V and V^* unchanged.

B. *Fixed Exchange Rates*

In the fixed exchange rates case we assume currencies are fully convertible at a parity pegged by the monetary authorities. In the absence of capital flows and sterilization policy, a trade imbalance is reflected in monetary flows. In the simplest metal money model, the world money supply is redistributed toward the surplus country at precisely the rate of the trade surplus. We assume that the world money supply is given and equal to $\overline{G}$, measured in terms of domestic currency. The rate of increase of the domestic quantity of money is therefore equal to the reduction in foreign money, valued at the fixed exchange rate $\bar{e}$:

$$(29) \qquad \dot{M} = -\bar{e}\dot{M}^*$$

where $\dot{M} \equiv dM/dt$.

For a fixed rate world we have to determine in addition to the real variables $\bar{\omega}$ and $\bar{z}$, the levels of money wages W and W^* as well as the equilibrium balance of payments associated with each short-run equilibrium. In the long run the balance of payments will be zero as money ends up

[11] This is a strong assumption since it makes spending independent of income and nonliquid assets even in the short run.

redistributed internationally to the point where income equals spending in each country. In the short run an initial misallocation of money balances implies a discrepancy between income and spending and an associated trade imbalance. To characterize the preferred rate of adjustment of cash balances in the simplest and most manageable way, we assume that spending by each country is proportional to money holdings.[12] On the further simplifying assumption that velocities are equal in each country, $V = V^*$,[13] world spending is equal to

$$VM + eV^*M^* \equiv V\bar{G} \tag{30}$$

For the tastes and technology specified in Section I, world spending on domestically produced goods is given by

$$V\bar{G}\int_0^{\tilde{z}} b(z)dz \equiv \vartheta(\omega)V\bar{G} \qquad \tilde{z} = A^{-1}(\omega) \tag{31}$$

In equilibrium, world spending on our goods must equal the value of our full-employment income WL:

$$WL = \vartheta(\omega)V\bar{G} \tag{32}$$

Equilibrium requires, too, that world spending on foreign goods equals the value of foreign full-employment income:

$$\bar{e}W^*L^* = [1 - \vartheta(\omega)]V\bar{G} \tag{33}$$

Equations (32) and (33) express what would seem to be the *joint* determination of real and monetary variables. But, in fact, we could have taken the shortcut of recognizing that the real equilibrium is precisely that of the barter analysis developed in Section I. Dividing (32) by (33) and substituting from (11) for the equilibrium relative wage $\bar{\omega}$, we can employ equations (32) and (33) to determine money wage levels.

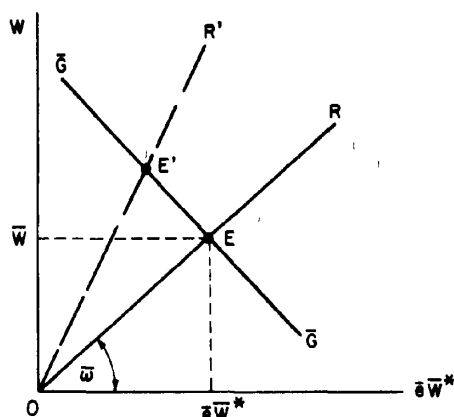



FIGURE 4

The equilibrium determined by equations (32) and (33) can be analyzed in terms of Figure 4. The figure emphasizes the separation of real and monetary aspects of the equilibrium under our assumptions of traded goods only, and no distribution effects. From the ratio of (32) and (33) we obtain the equilibrium relative wage $\bar{\omega}$ as a function of tastes and technology solely from the barter model. This equilibrium relative wage is plotted as the ray OR in Figure 4.

The equality of world income and spending

$$WL + \bar{e}W^*L^* = V\bar{G} \tag{30'}$$

is shown as the downward sloping straight line $\bar{G}\bar{G}$, which is drawn for given velocity, world quantity of money, and labor forces. Point E is the equilibrium where relative prices and the level of wages and prices are such that all markets clear. At a level of wages and prices higher than point E, there would be a world excess supply of goods, and conversely at points below E.

Figure 4 immediately shows some comparative static results. Thus a doubling of both countries' labor forces, from the analysis of the barter model, will leave the relative wage unaffected but will double

[12]The assumption that spending is proportional to cash balances is only one of a number of possible specifications. Conditions for this expenditure function to be optimal are derived in Dornbusch and Michael Mussa. In general, expenditure will depend on both income and cash balances.

[13]In the long-run equilibrium, higher V than V^* leaves us with a smaller share of the world money stock than foreigners, but with nominal and real income *shares* in the two countries the same as when $V = V^*$.

world output. Given unchanged nominal spending $V\overline{G}$, wages and prices will have to halve. This would be shown by a parallel shift of the $\overline{GG}$ schedule halfway toward the origin. A shift in demand toward the home country's output by contrast would rotate the OR ray to a position like OR' since it raises our relative wage. The ensuing monetary adjustment is then an increase in our money wage and money income and a decline in foreign wages, prices, and incomes (point E').

The real and nominal equilibrium at point E in Figure 4 is independent of the short- and long-run distribution of the world quantity of money. The independence of the real equilibrium derives from the uniform homothetic tastes. The independence of the nominal equilibrium is implied by identical velocities. What does, however, depend on the short-run distribution of world money is the transition periods' balance of payments. As in the absorption approach of Sidney Alexander (1952), we know this: when goods markets clear, the trade surplus or balance of payments $\dot{M}$ of the home country is equal to the excess of income over spending, or:

$$\dot{M} = \overline{W}L - VM \tag{34}$$

With the nominal wage independent of the distribution of world money, equation (34) therefore implies that the trade balance monotonically converges to equilibrium at a rate proportional to the discrepancy from long-run equilibrium:[14]

$$\dot{M} = V(\overline{M} - M); \overline{M} = \vartheta(\overline{\omega})\overline{G} \tag{34'}$$

The assumptions of this section were designed to render inoperative most of the traditional mechanisms discussed as part of the adjustment process: changes in the terms of trade, in home and/or foreign price levels, in relative prices of traded and nontraded goods (there being none of the latter), in double factoral terms of trade; and any discrepancies in the price of the same commodity between countries. The features of the adjustment process of this section rely on 1) identical, constant expenditure velocities, 2) uniform-homothetic demand, and 3) the absence of trade impediments. If velocities were constant but differed between countries, the absolute *levels* of money wages and prices, though not *relative* wages or prices, would depend on the world distribution of money. Relaxation of the uniform-homothetic taste assumption would make equilibrium relative prices a function of the distributions of spending. Finally, the presence of nontraded goods would, together with Ricardo's technology, provide valid justification for some of the behavior of relative prices and price levels frequently asserted in the literature; this behavior is studied in more detail in the next section.

C. *The Price-Specie Flow Mechanism under More General Conditions*

We now discuss the adjustment process to monetary disequilibrium and enquire into the price effects associated with a redistribution of the world money supply when there are nontraded goods. Common versions of the Hume price-specie flow mechanism usually involve the argument that in the adjustment process, prices decline along with the money stock in the deficit country, while both rise in the surplus country. There is usually, too, an implication that the deficit country's terms of trade will necessarily worsen in the adjustment process and indeed have to do so if the adjustment is to be successful.

Section IVB demonstrated that the redistribution of money associated with monetary imbalance need have no effects on real variables (production, terms of trade, etc.) and on nominal variables other than the money stock and spending. While this is

[14]Suppose $V > V^*$ and our share of the world money supply is initially larger than our equilibrium share. Then, as we lose M, total world nominal income and nominal GNP falls. Always our share of nominal world GNP stays the same under the strong demand assumptions. Total world real output never changes during the transition; only regional consumption shares change. Therefore, *both* countries' nominal price and wage levels fall in the transition, but such balanced changes have no real effects on either the transient or the final real equilibrium.

clearly a very special case, it does serve as a benchmark since it establishes that the monetary adjustment process would be effective even in a one-commodity world.

To approach the traditional view of the adjustment process more clearly and provide formal support for that view, we consider an extension to the monetary realm of our previous model involving nontraded goods. We return to the assumption that a fraction $(1 - k)$ of spending in each country falls on nontraded goods, and accordingly equations (32) and (33) become:

$$(32') \qquad WL = \vartheta(\omega)V\overline{G} + (1 - k)\gamma V\overline{G}; \quad \gamma \equiv M/\overline{G}$$

$$(33') \qquad \bar{e}W^*L^* = [k - \vartheta(\omega)]V\overline{G} + (1 - \gamma)(1 - k)V\overline{G}$$

These hold both in final equilibrium, and in transient equilibrium where specie is flowing. Equations (32′) and (33′) imply that the equilibrium relative wage does depend on the distribution of the world money supply. Solving these equations for the equilibrium relative wage we have:

$$(35) \qquad \overline{\omega} = \overline{\omega}(\gamma) \qquad \frac{\partial\overline{\omega}}{\partial\gamma} > 0$$

An increase in the home country's initial share in the world money supply γ raises our relative wage.

Using this extended framework, we can draw on the analysis of the transfer problem in Section II to examine the adjustment that follows an initial distribution of world money between the two countries that differs from the long-run equilibrium distribution.

Suppose our M is initially excessive, say from a gold discovery here. Assume also that the gold discovery occurred when the world was in long-run equilibrium with the previous world money stock. As a result of our excess M, we spend more than our earnings, incurring a balance-of-payments deficit equal to the rate at which our M is flowing out. In effect, the foreign economy is making us a real transfer to offset our deficit. As seen earlier, we, the deficit country, are devoting some of our excess spending to nontraded goods, shifting some of our resources to their production at the expense of our previous exports. We not only export fewer types of goods, but also import more types, and import more of each ($\overline{\omega}$ rises and $\bar{z}$ falls).

During the transition, while the real transfer corresponding to our deficit is taking place, our terms of trade are more favorable than in the long-run state. The new gold raises both their W^* and our W, but in addition, our W is up relative to their W^*. Therefore the price level of goods we continue to produce is up relative to the price level of goods they continue to produce. This is true both for our nontraded goods and for our exportables. The prices of goods we produce rise relative to the prices of goods they produce in proportion to the change in relative wages.

Thus the price levels in the two countries have been changed differentially by the specie flow and implied real transfer. But that does not mean that any traded good ever sells for different prices in two places. In fact the divergence in weighted average (consumer) price levels is due to nontraded goods. The price level will rise in the gold-discovering country relative to the other country the greater is the share of nontraded goods in expenditure, $1 - k$. It is a bit meaningless to say, "What accomplished the adjustment is the relative movements of price levels for nontraded goods in the two countries," since we have seen that the adjustment can and will be made even when there are no such nontraded goods. It is meaningful to say, "The fact that people want to direct some of their expenditure to nontraded goods makes it necessary for resources to shift in and out of them as a result of a real transfer, and such resource shifts take place only because the terms of trade (double-factoral and for traded goods) do shift in the indicated way."

The adjustment process to a monetary disturbance is stable in the sense that the system converges to a long-run equilibrium distribution of money with balanced trade. To appreciate that point, we supplement equations (32′) and (33′) with (34) that con-

tinues to describe the monetary adjustment process. We note, however, that now W and W^* are endogenous variables whose levels in the short run do depend on the distribution of the world money supply. A redistribution of money toward the home country would raise our spending and demand for goods, and reduce foreign spending and demand. As before, spending changes for traded goods offset each other precisely so that the net effect is an increase in demand for nontraded goods at home and a decline abroad. As a consequence our wages will rise and foreign wages decline. Therefore, starting from full equilibrium, a redistribution of money toward the home country will create a deficit equal to

$$(36) \quad d\dot{M}/dM = -V(1 - \delta) \quad 0 \leq \delta < 1$$

where δ is the elasticity of our nominal wages with respect to the quantity of money and is less than unity.[15] Equation (36) implies that the price-specie flow mechanism is stable.

It is interesting to observe in this context that the presence of nontraded goods in fact *slows down* the adjustment process by comparison with a world of only traded goods (contrary to J. Laurence Laughlin's turn of the century worries). As we saw before, with all goods freely tradeable, wages are independent of the distribution of money, and accordingly $\delta = 0$. Further we observe that the speed of adjustment depends on the relative size of countries. Thus the more equal countries are in terms of size, the slower tends to be the adjustment process.

In concluding this section we note that nontraded goods (and/or localized demand) are essential to the correctness of traditional insistence that the adjustment process necessarily entails absolute and relative price, wage, and income movements. They are, of course, in no way essential to the existence of a stable adjustment process, nor is there at any time a need for a discrepancy of prices of the same commodity across countries in either case.[16]

A final remark concerns the adjustment to real disturbances such as demand shifts or technical progress. It is certainly true that whether the exchange rate is fixed or flexible, real adjustment will have to take place and cannot be avoided by choice of an exchange rate regime. So long as wages and prices are flexible, it is quite false to think that fixed parities "put the whole economy through the wringer of adjustment" while in floating rate regimes "only the export and import industries have to make the real adjustment." It is true, however, that once we depart from flexible wages and prices there may well be a preference for one exchange rate regime over another. The next section is devoted to that question.

D. *Sticky Money Wages*

The last question we address in this section concerns the implications of sticky money wages. For a given world money supply, *downward* stickiness of money wages implies the possibility of unemployment. We assume upward flexibility in wages, once full employment is attained.

We start with a fixed exchange rate $\bar{e}$. The relation between wages and the world quantity of money is brought out in Figure 5. Denote employment levels in each country, as opposed to the labor force, by the new symbols $\bar{L}$ and $\bar{L}^*$, respectively; denote nominal incomes by Y and Y^*. The equality of world income and spending is again shown by the $\bar{G}\bar{G}$ schedule, the equation of which now is

$$(37) \quad V\bar{G} = Y + \bar{e}Y^* \equiv \bar{W}\bar{L} + \bar{e}\bar{W}^*\bar{L}^*$$

[15]The value of δ can be calculated from equations (32′) and (33′) to be

$$\delta = (1 - k)\frac{\gamma(1 - \gamma)}{\gamma(1 - \gamma) + \vartheta\epsilon}$$

where ϵ is the elasticity of the share of our traded goods in world spending, $\epsilon = -\vartheta'\omega/\vartheta > 0$. The elasticity δ is evaluated at the long-run equilibrium where $\gamma = \vartheta/k$. If $A'(z)$ falls slowly, ϵ will be large.

[16]The continuum Ricardian technology is special in that there can be no range of goods both imported and produced at home. Therefore, the cross elasticity of supply between nontraded goods and exports must be greater than the zero cross elasticity between nontraded goods and imports. Consequently, a transfer must shift the terms of trade (for goods and factors) in the stated orthodox way, favorably for the receiver.

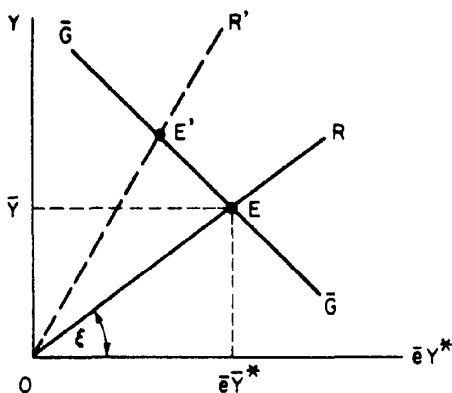



FIGURE 5

where $\overline{W}$ and $\overline{W}^*$ are the fixed money wages set at too high sticky levels. The schedule is drawn for given money wages, a given world quantity of money, and a pegged parity for $\bar{e}$. The ray OR now is predetermined by the given sticky relative wage $\bar{\omega} = \overline{W}/\bar{e}\overline{W}^*$. From equations (32) and (33) the ratio of money incomes $Y/\bar{e}Y^*$ is just a function of the relative wage now given exogenously by rigid money wages and the exchange rate:

$$(38) \quad Y/\bar{e}Y^* = \frac{\vartheta(\omega)}{1 - \vartheta(\omega)} \equiv \xi(\overline{W}/\bar{e}\overline{W}^*) \qquad \xi'(\omega) < 0$$

Point E is the nominal equilibrium where by assumption the world quantity of money is insufficient relative to wage rates to ensure full employment. Although that equilibrium is one with unemployed labor, it is efficient in other respects. Specifically, geographic specialization follows comparative advantage as laid out above, but now labor employed adjusts to sticky wage patterns of specialization.

Employment levels $\overline{L}$ and $\overline{L}^*$ now are determined by (39)

$$(39) \quad \overline{L} = \overline{Y}/\overline{W}, \quad \overline{L}^* = \bar{e}\overline{Y}^*/\overline{W}^*$$

where $\overline{Y}$ and $\overline{Y}^*$ are the equilibrium levels of nominal income determined by equations (37) and (38) or by point E in Figure 5.

Consider now the impact of a foreign increase in money wages. The effect of the implied reduction in our relative wages and the resulting increase in our relative income are shown in Figure 5 by the rotation from OR to OR'.

The new equilibrium is at E' where our money income and employment have risen while income and employment decline abroad. Thus an increase in the foreign wage rate, by moving the terms of trade against us, shifts comparative advantage and employment toward the home country. The extent to which the home country benefits from the adverse terms of trade shift in terms of employment will depend on both the substitutability in demand and the elasticity of the $A(z)$ schedule in Figure 1. We observe, too, that the move from E to E' will bring about a transitory balance-of-payments surplus. Given the initial distribution of money and hence of spending, the foreign decline in income and the increase at home implies that we will spend less than our income and therefore have a trade surplus. This surplus persists until money is redistributed to match the new levels of income at E'.

Next we move to flexible exchange rates. Under flexible rates an increase in the foreign money wage $\overline{W}^*$, given money supplies in each country, will similarly have real repercussion effects on relative prices and employment at home. Now employment in each country is determined by money supplies and prevailing wages:

$$(39') \quad \overline{L} = V\overline{M}/\overline{W}; \quad \overline{L}^* = V\overline{M}^*/\overline{W}^*$$

Given the employment levels thus determined, we know from the analysis of the earlier barter model that there is a unique relative wage at which the trade balance achieves equilibrium. The higher is $\overline{M}^*/\overline{W}^*$, the higher will be employment abroad—and, therefore, the higher will be our relative wage $\bar{\omega}$. It is thus apparent that an increase in the foreign money wage, $\overline{W}^*$, will reduce employment abroad. Employment declines only in proportion to the increase in wages and thus declines by less than it would under fixed exchange rates when specie is lost abroad.

We saw in the barter model that a reduc-

tion in effective foreign labor causes a decline in our relative wage, but that the decline in our relative wage falls proportionately short of the foreign reduction in labor. Now, at the initial exchange rate, the increase in foreign wages reduces our relative wage and their employment in the same proportion. The decline in our relative wage is therefore excessive. Domestic goods are underpriced and the exchange rate appreciates to *partly* offset the gain in cost competitiveness. The net effect is therefore a decline in our relative wage and an appreciation of our exchange rate (a decline in e) that falls short of the foreign increase in wages. Since our terms of trade unambiguously deteriorate without any compensating gain in employment, it must be true that welfare declines at home. Abroad, the loss in employment is offset by a gain in the terms of trade, but there too the net effect is a loss in welfare under our strong Mill-Ricardo assumption.

The adjustment to money wage disturbances under fixed and flexible rates differs in several respects. Under fixed rates employment effects are transmitted, while under flexible rates they are bottled up in the country initiating the disturbance. Under fixed rates the terms of trade move one for one with money wage, while under flexible rates exchange rate movements partly offset increases in the foreign money wage rate.

The difference between fixed and flexible rates in relation to the adjustment process is further brought out by an example of a real disturbance. Consider a shift in world demand toward our goods. Under fixed rates the resulting increase in our relative income will, from (38), move us in Figure 5 from E to E'. Employment rises at home and falls abroad. Demand shifts are fully reflected in employment changes. Under flexible rates, by contrast, with given wages and money, a demand shift has no impact on employment—as we observe from (39). At the initial exchange rate the demand shift would give rise to an excess demand for our goods and to an excess supply abroad. Domestic income and employment would tend to rise while falling abroad. The resulting trade surplus causes our exchange rate to appreciate until the initial employment levels and therefore trade balance equilibrium are restored. The demand shift is fully absorbed by a change in the terms of trade and a shift in competitive advantage that restores demand for foreign goods and labor.

Real and nominal equilibria are thus seen to be uniquely definable in our continuum model with constant-velocity spending determinants. The difference between sticky and flexible wage rates under fixed exchange rates is understandable as the difference between (a) having the crucial relative wage $\bar{\omega}$ be imposed in the sticky wage case with employments having then to adjust; or (b) having the full employments be imposed and $\bar{\omega}$ having to adjust. Under floating exchange rates, sticky nominal wages impose employment levels in each country and the crucial relative wage $\bar{\omega}$ then adjusts to those employment levels.

APPENDIX

Historical Remark

Figure 1 seems to be new. G. A. Elliot (1950) gives a somewhat different diagram, one that makes explicit the meaning of Marshall's 1879 "bales" (which, by the way, happen to work only in the two-country constant labor costs case). In terms of the present notations, Elliott plots for the *U.S* offer curve the following successive points traced out for all ω on the range $[0, \infty]$: on the vertical axis is plotted our total real imports valued in foreign labor units ("our demand for bales of their labor," so to speak), namely,

$$\int_{z}^{1} [P^*(z)/w^*]C(z)dz = \int_{z}^{1} a^*(z)C(z)dz$$

and on the horizontal axis, our total real exports valued in home labor units ("our supply of bales of labor to them"), namely,

$$\int_{0}^{z} [P(z)/w][Q(z) - C(z)]dz = L - \int_{0}^{z} a(z)C(z)dz$$

It is to be understood that $\tilde{z}$ is a function of ω, namely the inverse function $A^{-1}(\omega)$; also that $C(z)$ are the amounts demanded as a function of our real income L and of the $P(z)/W$ function defined for each, namely $\min[\omega a(z), a^*(z)]$. Because we have a continuum of goods, we avoid Elliott's branches of the offer curve that are segments of various rays through the origin. The reader will discern by symmetry considerations how the foreign offer curve is plotted in the same (L, L^*) quadrant, by varying ω to generate the respective coordinates

$$\left[\int_0^{\tilde{z}} a(z)C^*(z)dz, \quad L^* - \int_{\tilde{z}}^1 a^*(z)C^*(z)dz\right]$$

Our model forces the Elliott-Marshall diagram to generate a *unique* solution under uniform-homothetic demand. Unlike our Figure 1, the Elliott diagram can handle the general case of nonhomothetic demands in the two countries; but then, as is well known, multiple solutions are possible, some locally stable and some unstable. The price one pays for this generality is that, as Edgeworth observed, the Marshallian curves are the end products of much implicit theorizing, with much that is interesting having taken place offstage.

REFERENCES

S. Alexander, "The Effects of a Devaluation on the Trade Balance," *Int. Monet. Fund Staff Pap.*, Apr. 1952, *2*, 263–78.

J. S. Chipman, "A Survey of International Trade: Part I: The Classical Theory," *Econometrica*, July 1965, *33*, 477–519.

G. Debreu, "Economies with a Finite Set of Equilibria," *Econometrica*, May 1970, *38*, 387–92.

R. Dornbusch and M. Mussa, "Consumption, Real Balances and the Hoarding Function," *Int. Econ. Rev.*, June 1975, *16*, 415–21.

G. A. Elliott, "The Theory of International Values," *J. Polit. Econ.*, Feb. 1950, *58*, 16–29.

F. Graham, "The Theory of International Balances Re-Examined," *Quart. J. Econ.*, Nov. 1923, *38*, 54–86.

Gottfried Haberler, *The Theory of International Trade*, London 1937.

R. W. Jones, "Presumption and the Transfer Problem," *J. Int. Econ.*, Aug. 1975, *5*, 263–74.

James Laurence Laughlin, *Principles of Money*, New York 1903.

John S. Mill, *Principles of Political Economy*, London 1848.

David Ricardo, *On the Principles of Political Economy and Taxation*, 1817; edited by P. Sraffa, London 1951.

P. A. Samuelson, "The Transfer Problem and Transport Costs: The Terms of Trade When Impediments are Absent," *Econ. J.*, June 1952, *62*, 278–304; reprinted in Joseph Stiglitz, ed., *Collected Scientific Papers of Paul A. Samuelson*, Vol. 2, Cambridge, Mass., ch. 74.

———, "The Transfer Problem and the Transport Costs, II: Analysis of Effects of Trade Impediments," *Econ. J.*, June 1954, *64*, 264–89; reprinted in Joseph Stiglitz, ed., *Collected Scientific Papers of Paul A. Samuelson*, Vol. 2, Cambridge, Mass., ch. 75.

———, "Theoretical Notes on Trade Problems," *Rev. Econ. Statist.*, May 1964, *46*, 145–54; reprinted in Joseph Stiglitz, ed., *Collected Scientific Papers of Paul A. Samuelson*, Vol. 2, Cambridge, Mass., ch. 65.

S. Smale, "Structurally Stable Systems Are Not Dense," *Amer. J. Math.*, 1966, *88*, 491–96.

Frank W. Taussig, *International Trade*, New York 1927.

Jacob Viner, *Studies in the Theory of International Trade*, New York 1937.

C. Wilson, "On the General Structure of Ricardian Models with a Continuum of Goods: Applications to Growth, Tariff Theory and Technical Change," unpublished paper, Univ. Wisconsin-Madison, 1977.

Demand for International Media of Exchange

By K. ALEC CHRYSTAL*

"Just as every country needs a reserve of money for its home circulation so too it requires one for external circulation in the markets of the world."
Karl Marx

In the last two decades a great deal of attention has been directed at the determinants of the demand for money in a domestic economy context. A reasonable sketch of what has emerged from this enquiry is that people hold a stock of real money balances for the flow of transactions services that such stocks yield subject to a wealth, or portfolio constraint, and opportunity cost. As in reality a substantial proportion of all transactions take place across international frontiers, this reasoning might also suggest that nationals of one country should be expected to hold a portfolio of moneys which are accepted internationally, in addition to the domestic medium of exchange. This notion is not entirely new, as the above quote from Marx indicates, but in what follows the novel feature is an attempt to explain the aggregate composition of certain foreign currency balances as if the components were competing in the role of international media of exchange, or what have been called elsewhere "vehicle currencies." Section I will briefly outline some existing notions concerning international media of exchange, Section II will develop a model of the demand for such assets, and Section III will discuss the results.

*Lecturer in economics, Essex University. This paper draws heavily on my doctoral thesis presented at the University of Essex. The help and encouragement of Frank Brechling, David Laidler, and Christopher Bliss is gratefully acknowledged. James Alt and Padmini Kurukulaaratchy have given computational assistance. The final version has benefited from discussions with John Williamson and comments from George Borts and an anonymous referee.

I

There are three separate strands to the existing arguments concerning international media of exchange. First, the microeconomic analysis of transactions demand in an international context; secondly the "aggregation" proposition that there are significant economies resulting from the conduct of trade in terms of few currencies; and thirdly, the need to determine which currencies actually do get used if the choice is largely left to market forces.

Alexander Swoboda has applied the Baumol transactions demand for cash inventory model to foreign currency balances in a direct way. He postulates that an individual has a known steady stream of foreign currency payments to make and withdraws such cash from domestic interest bearing assets in discrete lumps. Here the familiar result must emerge that foreign currency balances will on average be proportional to the square root of transactions made in that currency. Thus, so long as a specific currency can be uniquely associated with a specific set of transactions, there is nothing in principle different about foreign currency transactions. The model developed below, however, permits internationally acceptable moneys to compete for the role of financing trade, so that the relationship between specific currencies and transactions is less direct.

The square root rule is used by Swoboda to make the second point, which is that in a multicurrency area world it will not be the case that intercurrency area trade is conducted in terms of all the different national currencies according to the origins or destination of the trade. The resources that traders would have to devote to holding cash balances would be much greater if they had a multitude of transactions requiring different currencies, than if all international

transactions could be conducted in one currency. This currency would then be what Swoboda calls the vehicle currency, or what is here called the international medium of exchange, or simply international money. An alternative path to the same conclusion is that offered by Karl Brunner and Allan Meltzer who, by analyzing the informational efficiency of monetary exchange, "... suggest by implication the benefits that would accrue to the world economy from the use of a medium of exchange" (p. 804). Their key proposition is that the marginal cost of acquiring information about the properties of any asset declines with the frequency with which that asset is used. And Ronald McKinnon uses both arguments to suggest that "... private traders would concentrate their transactions in the most suitable major currency in order to economise on inventory-carrying costs and to minimise the informational uncertainty arising from floating rates" (p. 14). Swoboda makes a further important point that where there are risks involved, specialization in the use of a single international money may not be complete since there will be gains from diversification.

The final question is what factors will determine which of the existing currencies come to be used as international money. Swoboda is again of assistance in suggesting that "... asset-exchange costs play an important role in this choice.... It is likely that asset exchange costs depend inversely on the size of the market for a particular asset.... The size of the market for a particular currency depends, in turn, in part on the size of a country's foreign transactions and, therefore, on the volume of its external trade ..." (p. 10). Secondly, since holders are likely to be risk averse, the domestic market of the currency chosen should be characterised by "depth, breadth and resilience" since there is a greater probability of loss from selling on a small market than on a large one. Finally, for similar reasons, no currency, the exchange value of which is likely to fluctuate widely, would be held as an international money for very long.

In summary, then, while it can be suggested that the conventional transactions model is not entirely appropriate for international moneys, since currencies are not uniquely tied to a given set of transactions (indeed this variable relation between transactions and currency is the most interesting extra dimension of the international money demand problem), it does lend support to the argument that only a few currencies will actually serve as international media. The currencies which do come to be used are likely to be those of the dominant trading nations and/or those with well-developed domestic money markets, though in practice these would appear to be the same.

II

It is now presumed that a small number of currencies are to be found circulating at large in the international economy, playing the role of transactions media. The problem to be addressed is that of explaining the composition of these balances when viewed as elements of a portfolio. It is convenient to theorize first at the level of a "typical" individual who has a total stock of these currencies given to him at the beginning of each period of time as a result of the outcome of a wider economic process, and who then arranges this portfolio in what he may consider to be an optimal manner. The possibility of rearranging the portfolio is then absent until the beginning of the next period, and the structure of the portfolio is assumed independent of his wider asset position. It is then possible to derive a set of demand functions for the currencies which has the following form:[1]

$$x = Ka + Kr + hW \quad (1)$$

where x is the vector of currencies in the

[1]The holder maximizes the expected utility of the portfolio. Utility is an exponential function of real income, and income is the sum of the pecuniary and nonpecuniary services of each currency. The two types of service flow are assumed to be dimensionally equivalent. The marginal service flow is assumed constant.

portfolio, a is the vector of expected nonpecuniary service flows yielded by each currency, r is the vector of pecuniary yields, W is the portfolio size, K is a matrix of parameters which is symmetric with zero row and column sums, and h is a vector with elements summing to unity. The key variables in the equation are obviously the vectors a and r, and it is therefore necessary to give them further thought before attempting to estimate the parameters.

The conceptual experiment under consideration is concerned with the allocation of a portfolio for a succession of discrete periods. Currencies are measured by their current dollar value. The pecuniary yield will have two separate elements, one a coupon interest payment, the other a change of exchange rate. Expressing exchange rates in currency units per dollar, the pecuniary yield on currency i is

$$(2) \qquad \frac{SR_t^i}{SR_{t+1}^i}(1 + r_t^i) - 1$$

where SR_t^i and SR_{t+1}^i are the spot rates at the beginning and end of the period, respectively, and r_t^i is the coupon yield on the ith currency over the period. As an empirical matter r_t^i is assumed to be the rate ruling at the beginning of each period, but exchange rate expectations are normally more troublesome. The assumption used below is that exchange rates are expected to fluctuate randomly about their current level so that the rate expected at the end of the period is that ruling at the beginning.[2] This is a reasonable assumption for the data period which is broadly the 1960's.

The nonpecuniary yield corresponds to what are commonly called the "liquidity" or "transactions" services of money, which are considered to be valuable in the sense that there is some interest rate equivalent which would be forfeited to get them. A proxy for the services of each currency can be derived from the asset-exchange cost argument of Swoboda, referred to in Section I above. Given that there is a small group of currencies available to perform the role of international money, then the nonpecuniary services of each will be greater, *ceteris paribus*, the more important is the country of issue in world trade. The importance of each country in world trade is proxied by the value of its exports. The functional relationship between trade shares and services is assumed to be linear

$$(3) \qquad a = VT$$

where a is the vector of unit service flows, V is a diagonal matrix of parameters, and T is a vector of export values from the "banker" countries to the typical holder. So (1) may now be rewritten

$$(4) \qquad x = K\,VT + Kr + h\,W$$

The approach of methodological individualism which has been adopted by basing behavioral reasoning on a "typical" individual presumes the aggregate function, since to be typical one must be representative of the behavior of all. Thus to move from the typical individual to the aggregate it is merely necessary to sum over all such individuals. The system of demands by the jth individual is

$$(5) \qquad x_j = K\,VT_j + Kr_j + h\,W_j$$

and the demand for the ith currency by the jth individual can be written (where four currencies are assumed as below)

$$(6) \quad x_j^i = b_1^i T_{1j} + \ldots + b_4^i T_{4j} + \ldots + k_1^i r_{1j} + \ldots + k_4^i r_{4j} + h^i W_j$$

where $b_1^i = k_1^i v_{11}$

Adding across individuals

$$(7) \quad \sum_{j=1}^{n} x_j^i = b_1^i \sum_j T_{1j} + \ldots + b_4^i \sum_j T_{4j} + \ldots + k_1^i \sum_j r_{1j} + \ldots + k_4^i \sum_j r_{4j} + h^i \sum_j W_j$$

$\Sigma_j T_{ij}$ is simply the sum of trade with each individual, that is, total export value, and $\Sigma_j W_j$ is the sum of the currency portfolios

[2] Two other assumptions are reported in my dissertation. The use of the *ex post* exchange rate changes did no better and covered yields did much worse.

of each individual, that is; total currency balances. Since all are assumed to calculate the pecuniary yield in dollars, $\Sigma_j r_{ij}$ is simply $n\, r_i$ where there are n such individuals, so that the matrix of pecuniary yield parameters will differ from that of the nonpecuniary yield by a factor of n.

To say no more than this would, however, be to ignore all of the most interesting aspects of this problem, since the data to be used below include both private and official holdings. At first sight many would regard such an aggregate treatment as erroneous, especially since transactions services are explicitly relevant and yet official bodies hardly trade at all. The counter argument, however, is that there is considerable interdependence between the currency holding of the trading sector of an economy and those of its central bank. An argument is developed in my dissertation that, to some extent, central bank reserves can be thought of as being held on behalf of the trading sector, but there is also an interdependence as to composition. Consider there to be some given total of foreign money held by a particular country which is divided in some proportion between private and official holdings. If some trader is regularly trading in terms of some currency and his central bank does not hold balances of it upon which he can draw, then on average he will have to hold a balance of it for himself. Similarly, if traders do hold sufficient stocks such that they never need to enter the foreign exchange market, then the central bank would have no need for balances of that particular currency. In reality the picture is complicated by restrictions such as exchange controls, but that particular example increases the interdependence by making reserve pooling mandatory.

Thus, while it is conceded that the aggregate of reserves held officially may differ considerably from that which would be jointly demanded by traders when the composition of actual official and private holdings is looked at, the two are not independent. So it might be expected that the model developed above should apply at least as well to total balance composition as it would to either private or official balances separately (though data limitations make a proper disaggregated treatment difficult anyway).[3]

Some lesser problems must now be discussed, namely Eurocurrency, gold, banker country holdings, and identification. The importance of Eurocurrencies is central to the question of the correct measure of currency balances. Should dollar balances, for example, be measured as *U.S.* short-term liabilities to foreigners as perceived in New York, or should Eurodollar deposits be added in some way. The view taken here is that Euromarkets are simply a system of secondary financial intermediation which do not add to the money stock and are therefore excluded, but to the extent that there is in fact net credit creation, bias will be introduced. To include aggregate Eurocurrency data as it stands, however, would introduce greater error since there would be a large element of double counting; the interbank transactions are not totally netted out (at least for the period studied); some Eurodeposits are too long-term for inclusion; and finally, some Eurodeposits may be held by domestics of the country of issue for domestic reasons other than as an international medium of exchange.

Gold is also excluded since it was not judged to be an effective trading currency; private holdings are unknown; supply conditions are entirely different from those of the other currencies; and gold is a commodity which may be demanded for entirely nonmonetary purposes. Implicit in the exclusion of gold is the assumption that although gold may be a substitute for foreign exchange in general, its effect can be subsumed within the overall portfolio constraint and will not affect the composition of foreign exchange portfolios of given size.

A further problem arises from the fact that choice between a number of currencies is studied as if the typical holder potentially holds some of each. There are two separate

[3]Some limited disaggregated evidence is available in my dissertation. This indicates that the sterling equation may suffer from aggregation errors.

aspects to this. First, private traders of banker countries do in fact hold their own national money and may use it for external trade, though this will not be included in the data used, rather it would appear in the domestic money supply. Secondly, central banks of banker countries only have a reserve asset choice among currencies other than their own (or may be considered to run a portfolio of assets and liabilities). Indeed the United States is often thought to have no effective official foreign portfolio at all, though in fact the portfolio did become quite sizeable during the period of study, subsequently to fall. Notionally, therefore, the official reserves of banker countries should be treated separately, since they are operating under different constraints than those postulated. Those elements of the domestic moneys used internationally should be included since if they were not being used some other international money would be. However, neither the exclusion nor the inclusion is possible given present data availability, so the best that can be hoped for is that in the aggregate, the biases thereby induced will be small and to some extent offsetting.

The position of banker countries might appear to be of some concern also for the treatment of speculation within the present framework. There are three separate ways in which speculation could affect the model. First, there are the relative expected values of the international moneys themselves, and this is incorporated as it affects the pecuniary yields. Second, there is speculation involving the expected value of other currencies, and this would have influence only upon the overall portfolio constraint, not upon its composition. But, thirdly, the possibility of speculation against a banker country could affect its own reserve holdings and, in so far as this is likely to affect some more than others, biases could result.

The final question to be discussed in this section is that of identification. The model developed above leads to the derivation of demand functions on the part of foreigners for holdings of balances of particular currencies, and it is necessary to ascertain whether an estimation of these as they stand will indeed provide valid estimates in principle. This is clearly entirely dependent upon what supply conditions are believed to have been. The data period taken in this study corresponds roughly to the decade of the 1960's, which was a time of almost complete fixity of exchange rates and of convertibility (at least for the central currencies concerned). Convertibility is taken to mean that the central bank of the country of issue of a currency agrees to exchange that currency into other currencies as required, at or close to a fixed price. In such circumstances, foreigners as a whole need not hold on to a particular currency if they do not wish to, since by definition it can be freely changed into other currencies. This means that foreign held balances in conditions of convertibility are necessarily demand determined. Complete convertibility of the currencies involved does then represent a sufficient condition for identification of the demand equations; though a weaker set of conditions is sufficient for what follows (except where world trade is included). These conditions may be called "weak convertibility" and would pertain so long as there is sufficient convertibility to ensure that the composition of currency portfolios is demand determined even if the balances of, say, one particular currency are exogenously given. Weak convertibility may be defined as convertibility of all except one currency. Thus, if, say, the dollar were inconvertible, or holders behaved as if it were so, such that foreigners could or would not change dollars into foreign exchange in New York, the foreign held dollar stock would be supply determined. But so long as the issuers of substitute currencies were prepared to accept dollars in exchange for their own currency, then the proportion of dollars in currency portfolios remains demand determined.[4] Formally, then, it can be said that weak convertibility is generally assumed to be identifying, and the imposition of the set of portfolio restrictions is overidentifying.

[4]Subject to the banker country exclusion discussed above.

III

The model as it stands in equation (4) says that the exogenously given portfolio W will be distributed among the currencies x, according to the relative pecuniary return on each, r, and the relative importance of the country of issue in world trade T (where T is a vector of export values), r and T being assumed exogenous to holders. All parameters, K, V, and h are identified, but to estimate V directly requires a non-linear estimation procedure. Such estimates are reported below, but first, it was convenient to perform most estimates on the linearized form.

$$x = BT + Kr + hW \tag{8}$$

where $B = KV$ and it is assumed that the restrictions applicable to K also apply to B, that is, the diagonal elements of V are identical.

The linear estimation procedure may be outlined as follows. Let

$$y_i = X_i \beta_i + e_i \tag{9}$$

be the ith equation of an m equation regression system, where y_i is an $n \times 1$ vector of observations on the ith currency, X_i is an $n \times j$ matrix of observations on the j independent variables; β_i is a $j \times 1$ vector of coefficients and e_i is an $n \times 1$ vector of disturbances. The complete system may be stacked and written

$$Y = XB + E \tag{10}$$

Then the ordinary least squares (OLS) estimate of B is

$$\hat{B} = (XX)^{-1}X'Y \tag{11}$$

and the restricted least squares (RLS) estimate is

$$\tilde{B} = \hat{B} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{B}) \tag{12}$$

where the a priori restrictions can be expressed

$$r = RB \tag{13}$$

The estimates actually performed used the University of Essex MOARE program which uses a two-stage RLS procedure similar to that above but with the additional feature that estimates of the errors from the first stage are used to calculate the second stage, that is,

$$\tilde{B} = \hat{B} + (X'\Omega^{-1}X)^{-1}R' \cdot [R(X'\Omega^{-1}X)^{-1}R']^{-1}(r - R\hat{B}) \tag{14}$$

where Ω is the, appropriately stacked, estimate of the contemporaneous error variance-covariance matrix. The non-linear estimates were performed on the University of Essex BMDX95 program, which uses a Gauss-Newton iterative procedure.

Estimates were performed taking four foreign held currency balances (pound sterling, dollar, French franc, and deutschemark) quarterly for the period 1961 first quarter to 1970 second quarter (38 observations). The balances data correspond where possible to short-term liquid liabilities to foreigners, which include both bank deposits and marketable short-term government debt. Interest rates chosen were three-month Treasury Bill rates for the United Kingdom and United States, and three-month call money rates for France and West Germany. A dummy variable for the first 8 observations reflects the change in sterling series, and it is included in all equations to preserve the portfolio symmetry.

Table 1 reflects the estimates of equation (8) where all a priori restrictions are imposed and where variables have been transformed to correct for autocorrelation.[5] The overwhelming positive feature is that the sign pattern expected is strongly evident. The matrix of interest coefficients has positive diagonals and dominantly negative off-diagonals and the only interest complementarity is between the franc and deutschemark which is not a counterintuitive result. The matrix of trade responses is also well determined, though there is a significant wrong sign on French exports in the franc equation. One dominant feature which is a recurrent result in all the estimates per-

[5] An additional stage of estimation was added taking an estimate of the autocorrelation coefficient from the first stage.

TABLE 1—LINEAR MODEL WITH AUTOCORRELATION TRANSFORM

	Sterling	Dollar	Franc	*DM*
U.K. TBR	552	−212	−75	−264
	(3.4)	(1.6)	(1.2)	(4.1)
U.S. TBR	−212	383	−166	−4.5
	(1.6)	(1.94)	(1.5)	(.07)
French *CMR*	−75	−166	137	104
	(1.2)	(1.5)	(1.8)	(3.0)
German *CMR*	−264	−4.5	104	164
	(4.1)	(.07)	(3.0)	(5.1)
U.K. Exports	1 7	−.66	−.06	−.99
	(4.4)	(2.8)	(.34)	(5.9)
U.S. Exports	−.66	.37	.26	.02
	(2.8)	(1.9)	(2.4)	(0.2)
French Exports	−.06	.26	−.74	.54
	(.34)	(2.4)	(3.3)	(3.7)
German Exports	−.99	.02	.54	.51
	(5.9)	(0.2)	(3.7)	(4.58)
Total Stock	.36	.6	.03	.03
	(1.3)	(28)	(1.8)	(2.4)
Data Dummy	−474	−805	1114	329
	(.88)	(2.1)	(5.5)	(1.4)
R^2	−.4	.98	.6	.85
D.W.	1.7	1.6	1.6	1.7

Notes: Variables are measured in current dollar values; *t*-statistics shown in parentheses; *D W.* is the Durbin-Watson statistic, number of observations is 38, Period 1961I to 1970II

TBR = Treasury Bill rate (3-month)
CMR = Call Money rate (3-month)
Total Stock = Sum of dependent variables
Dollar = *U.S.* short-term liquid liabilities to foreigners; source: Federal Reserve
Sterling = Nonresident sterling balances; source: Bank of England
Franc = External public debt plus foreign deposits with Bank of France; source: Institut National
DM = Short-term liabilities of German banks to foreigners; source: Deutschebundesbank
Data Dummy = 1961-I-1962-IV
Interest rates and exports; source: International Monetary Fund

formed is the strong substitutability of sterling and deutschemarks as evidenced by the relevant coefficients in both the interest and the trade matrices.

There is only one clearly unsatisfactory feature of the results in Table 1 and that is the poor R^2 in the sterling equation. It is obvious that the proximate determinant of this being negative is the absence of an intercept, but a number of other features are worthy of brief comment. In my dissertation it is demonstrated that an equation identical to that in Table 1 works acceptably well (i.e., with an R^2 in excess of .8) if private and official holdings are estimated separately. Below it is also demonstrated that, even in the absence of an intercept and in the aggregate, the sterling equation is well determined so long as lagged adjustment is allowed. And it is also the case that the effects of structural or institutional changes need to be taken into account. One such change was the Basle Agreement which became effective on September 25, 1968 under which the *U.K.* undertook to guarantee the dollar value of the bulk of sterling reserves in exchange for a major support facility.[6] An intercept dummy was introduced to capture the effects of this agreement and the result as presented in

[6]See Bank of England, p. 170.

TABLE 2—LINEAR MODEL WITH BASLE DUMMY

	Sterling	Dollar	Franc	DM
U.K. TBR	431	−218	6.4	−219
	(2.9)	(1.6)	(1.2)	(3.9)
U S TBR	−218	480	−233	−28.3
	(1.6)	(2.2)	(2.2)	(.47)
French CMR	6.4	−233	198	30.9
	(1.2)	(2.2)	(2.5)	(1.0)
German CMR	−219	−28.3	30.9	217
	(3.9)	(.47)	(1.0)	(6.8)
U.K. Exports	1.3	−.73	.14	−.69
	(3.2)	(2.7)	(.94)	(3.8)
U.S. Exports	−.73	.4	.09	.23
	(2.7)	(1.5)	(.8)	(1.8)
French Exports	.14	.09	− 17	−.06
	(.94)	(.8)	(.95)	(.45)
German Exports	−.69	.23	−.06	.5
	(3.8)	(1.8)	(.45)	(3.4)
Total Stock	.38	.58	.02	.02
	(13.6)	(22 3)	(1.4)	(1.3)
Data Dummy	−894	−587	1050	415
	(2.2)	(1.7)	(7.6)	(2.4)
Basle Dummy	−3390	1368	653	1317
	(6.6)	(3 2)	(2.2)	(4.7)
R^2	.44	.99	.86	.93
D W.	1.3	1 2	1.4	1.4

Notes: See Table 1; Basle Dummy = 1968-III-1970-II

Table 2. It is clear that the Basle Dummy has considerable explanatory power. There was a shift out of sterling in excess of $3 billion which indicates that even the Basle Agreement was not sufficient to compensate for the qualitative decline in sterling which made support necessary. The positive feature of the sign pattern of coefficients is somewhat strengthened though it should be noted that no further attempt was made to correct for autocorrelation.

When judged by the R^2, it is clear that the dollar equation has so far demonstrated the best overall goodness of fit. It might be objected, however, that since dollars are in excess of two thirds of total balances, the inclusion of the total in the dollar equation is bound to provide a good explanation. To meet this objection it is necessary to introduce the auxiliary hypothesis that the total stock of currencies held, W, is itself demand determined, and that this stock demand depends solely on the value of world trade which is exogenous. This can be thought of as a two-part decision hypothesis; in the first instance the aggregate value of international moneys demanded depends upon the value of world trade, yet the composition of the aggregate still depends upon the relative pecuniary and nonpecuniary services of the components. So (8) can now be written:

$$x = a + BT + Kr + m\overline{T} \quad (15)$$

where it is assumed that

$$W = d + y\overline{T} \quad (16)$$

so $a = hd$ and $m = hy$; $\overline{T}$ is the value of world imports so it is being assumed that there is a linear relationship between W and $\overline{T}$.

The result of estimating equation (15) is presented in Table 3. It is clear that world trade works very well and that the general characteristics of the sign patterns of coefficients are retained. Only the franc shows no significant responses to trade in either the aggregate or composition so it is reasonable to conclude that the franc is not a significant international medium of exchange. The dollar is sensitive to trade composition where only the portfolio constraint is in-

TABLE 3—LINEAR MODEL, WORLD TRADE

	Sterling	Dollar	Franc	DM
U.K. TBR	217	−59	22.8	−181
	(2.1)	(.06)	(.05)	(4.02)
U.S. TBR	−59	455	−339	−57
	(.06)	(2.85)	(4.3)	(1.3)
French *CMR*	22.8	−339	278	38
	(.05)	(4.3)	(5.4)	(1.5)
German *CMR*	−181	−57	38	200
	(4.02)	(1.3)	(1.5)	(6.9)
U.K. Exports	0.7	−.26	.04	−.5
	(1.5)	(1.0)	(.2)	(2.1)
U.S. Exports	−.26	.31	.03	−.07
	(1.0)	(1.4)	(.24)	(.63)
French Exports	.04	.03	−.02	.01
	(.2)	(.24)	(.9)	(.07)
German Exports	−.5	−.07	.01	.5
	(2.1)	(.63)	(.07)	(2.6)
World Trade	9.6	54.7	2.06	8.4
	(2.55)	(17.6)	(1.3)	(4.1)
Data Dummy	−1593	1656	1126	909
	(4.6)	(2.6)	(6.8)	(4.3)
Constant	9061	−596	−241	−174
	(8.43)	(.5)	(.05)	(2.7)
R^2	.76	.94	.79	.89
D.W.	.91	1.2	1.1	.8

Notes: See Table 1; World Trade = world imports; source: United Nations

cluded, but its position is obviously dominated by the effects of world trade in general. Sterling and the deutschemark remain responsive to both the trade aggregate and composition. The interest matrix also retains its desired pattern with such strength that the continued presence of autocorrelated residuals should not cause excessive alarm.

The final estimates to be presented are those of the parameters of equation (4) which are non-linear. However, an additional feature is introduced here, namely that the observed holdings are assumed to be only partially adjusted towards some desired level. The equation to be estimated is then

$$x_t = MKVT_t + MKr_t + MhW_t + (I - M)x_{t-1} \tag{17}$$

where M is a matrix of adjustment parameters. For present purposes M is assumed to be diagonal but it should be noted that this violates portfolio consistency since it is not possible for one element of a portfolio to be in disequilibrium alone. The results should therefore be taken as indicative rather than exact.

The results appear in the Appendix. The parameter estimates are of the relevant elements of the corresponding matrices so that, for example, m_{11} to m_{44} are the diagonal elements of M, being the adjustment coefficients in the sterling, dollar, franc, and deutschemark equations, respectively. The parameters k_{11}, etc. are those of the service yield matrix K, where the fourth row and column are implied by the others and so are not estimated. The elements of V are the trade response parameters which have previously been unidentified, and it is of interest to note that the size ordering of these is exactly the ranking one would expect to attach to these currencies as media of exchange, that is, dollar, sterling, deutschemark, franc. The general characteristics of the overall equation are satisfactory. The sign pattern of K is as required and the R^2 of each equation, even sterling, is reason-

TABLE 4—F-RATIO TEST FOR ACCEPTABILITY OF PORTFOLIO RESTRICTIONS

Table 1	Linear Model (12, 84)	= 4.55
Not Reported	Linear Model + Intercept (12, 81)	= 2.4[a]
Table 2	Linear Model + Basle Dummy (12, 81)	= 1.87[b]
Not Reported	Linear Model + Basle Dummy + Intercept (12, 78)	= 1.603[b]

Note: The degrees of freedom of the numerator and denominator are shown in parentheses.

[a]The test statistic exceeds the 95 percent acceptable level but not the 99 percent.

[b]Implies acceptability at the 95 percent level.

able. The adjustment coefficients cannot be accepted literally, but they do indicate a wide discrepancy of adjustment speeds with the dollar appearing 70 percent adjusted in one quarter on the one extreme and sterling being 7 percent adjusted on the other. The adjustment process obviously requires further investigation.

An *F*-ratio test[7] was performed for acceptability of the portfolio restrictions in the linear model. Test statistics for Tables 1 and 2 are presented in Table 4. These show that the restrictions on Table 1 are not acceptable but the inclusion of an intercept would make them acceptable at the 99 percent level. Inclusion of a Basle Dummy, however, makes the restrictions acceptable at the 95 percent level, with or without an intercept.

IV. Summary and Conclusions

This study takes as its starting point the observation that substantial foreign held balances of a small number of currencies exist, and attempts to explain changes in the structure of these balances "as if" they were being demanded as international media of exchange. In particular, attention is directed at dollars, pounds sterling, French francs, and deutschemark balances, and the proportion held is deemed to depend on the relative pecuniary and nonpecuniary returns, the latter being related to the importance of the country of issue in world trade.

The main empirical conclusion is that the pattern hypothesized is broadly supported. Interest rate effects are well established and robust, and for the study period expected exchange rate adjustments do not appear to seriously hinder estimation, though a dummy for the period since the Basle Agreement makes a significant explanatory contribution. The trade structure effects are generally less well established, though substitutability between sterling and deutschemarks is strongly evident. Dollar balances appear to be dominantly influenced by the aggregate level of world trade, rather than its composition, and the franc does not appear to be influenced by trade at all, thus bringing into question its role as a true international medium of exchange. The portfolio restrictions imposed upon the linear model are found to be acceptable at normal probability levels when the model is extended to include a dummy for the Basle Agreement.

The approach taken, although just about the simplest possible in this context, has two features of wider interest. First the notion of transactions demand is seen in a new light. There is a choice of moneys, they are not uniquely associated with specific transactions, and yet relevant transactions can be observed. This contrasts with the "domestic" demand for money literature where the money is presumed to be unique and the "work done" by money tends to be proxied by income or output rather than by transactions themselves. Secondly, changes in holdings of the balances studied have often been looked at as part of capital flows for the banker countries. This approach is complementary to the task of explaining such flows. Finally, it would seem that the empirical work of which this is but a modest be-

[7]See Henri Theil, p. 314.

ginning is a necessary precondition for assessing the required growth and composition of international money balances.

APPENDIX

Non-Linear Model with Partial Adjustment

The estimated equation system is $x = MKVT + MKr + MhW + MdD + (I - M)x_{-1} + e$

Variables

x_1 = sterling; x_2 = dollars; x_3 = francs; x_4 = deutschemarks

T_1 = *U.K.* exports; T_2 = *U.S.* exports; T_3 = French exports; T_4 = German exports

r_1 = *U.K. TBR*; r_2 = *U.S. TBR*; r_3 = French *CMR*; r_4 = German *CMR*

$W = \sum_i x_i$; D = data dummy 1961-I–1962-IV; x_{-1} = one-period lagged dependent variables

Parameter Estimates (these are elements of the corresponding matrices); asymptotic standard errors are shown in parentheses; and the asterisk indicates significance at the 95 percent level.

m_{11} = .072; m_{22} = .7*; m_{33} = .33*;
(.051) (.08) (.13)

m_{44} = .09
(.07)

k_{11} = 919*; $k_{12} = k_{21}$ = −195*; k_{13} =
(381) (63.2)

k_{31} = −152*; k_{22} = 17.3; $k_{23} = k_{32}$ = 55;
(30.1) (20.2) (44)

k_{33} = 279; $k_{14} = -(k_{11} + k_{12} + k_{13})$, etc.
(164)

v_{11} = .015*; v_{22} = .02; v_{33} = .0015;
(.005) (.018) (.0033)

v_{44} = .004*
(.0019)

h_1 = .085; h_2 = .75*; h_3 = .09;
(.24) (.05) (.086)

$h_4 = 1.0 - (h_1 + h_2 + h_3)$

d_1 = −139; d_2 = −753*; d_3 = 698*;
(157) (193) (208)

$d_4 = -(d_1 + d_2 + d_3)$

Equations

1) Sterling R^2 = .85, *D.W.* = 2.1
2) Dollar R^2 = .993, *D.W.* = 1.6
3) Franc R^2 = .84, *D.W.* = 1.5
4) Deutschemark R^2 = .95, *D.W.* = 1.4

REFERENCES

W. Baumol, "The Transactions Demand for Cash," *Quart. J. Econ.*, Nov. 1952, *66*, 545–56.

K. Brunner and A. Meltzer, "The Uses of Money," *Amer. Econ. Rev.*, Dec. 1971, *61*, 784–805.

K. A. Chrystal, "Demand for International Media of Exchange," unpublished doctoral dissertation, Univ. of Essex 1975.

Karl Marx, *Das Kapital*, London 1920.

Ronald I. McKinnon, *Private and Official International Money: The Case for the Dollar*, in *Essays in International Finance*, Princeton University, No. 74, 1969.

Alexander K. Swoboda, *Eurodollars, a Suggested Interpretation*, in *Essays in International Finance*, Princeton University, No. 64, 1968.

Henri Theil, *Principles of Econometrics*, New York 1971.

Bank of England, *Quarterly Bulletin*, June 1974, *14*.

Board of Governors of the Federal Reserve System, *Fed. Res. Bull.*, Washington, various issues.

Deutschebundesbank, *Monthly Report*, Frankfurt, various issues.

Institut National de Statistique et des Etudes Economique, *Bulletin Mensuel*, Paris, various issues.

International Monetary Fund, *International Financial Statistics*, various issues.

United Nations, *Mon. Bull. Statist.*, New York, various issues.

Structural Expectations and the Effectiveness of Government Policy in a Short-Run Macroeconomic Model

By STEPHEN J. TURNOVSKY*

Over the past few years, expectations, and in particular inflationary expectations, have come to play a central role in macroeconomic theory. In modeling these expectations two alternative procedures have typically been adopted. One approach is to specify them by some autoregressive function of the variable being predicted, so that at any specified time t, say, they can be treated as given, being predetermined by past values of that variable. The most common of such autoregressive procedures is the adaptive expectations hypothesis in which the forecast is adjusted in proportion to the immediate past forecast error.

Despite their widespread use in a variety of contexts, these autoregressive hypotheses have periodically come under severe criticism, especially when applied to predicting *endogenous* variables. The objections have been along the following lines. By forecasting an endogenous variable using past values of that variable alone, one is clearly disregarding a considerable volume of available information relevant to that variable. In particular one is ignoring any knowledge one might have of the economic structure being analyzed, the very purpose of which is to provide predictions of the endogenous variable. Indeed there is no reason for the predictions generated from the autoregressive scheme to be consistent with those implied by the model. Hence it has been argued that if forecasters are aware of the structure of the relevant economic system, the rational way for them to form their expectations is to base them on the predictions of the economic model. Of course this criticism does not apply to predictions of exogenous variables, since by definition these are not explained within the framework of the model.

This hypothesis, known as the rational expectations hypothesis, originated with John Muth. Formally it requires the forecaster's predicted value for period t, say, to equal the expected value of that variable as predicted by the system, conditional on all information available at the time the prediction is made (usually time $(t - 1)$).

The insistence that expectations be rational is also open to objections. In order for a prediction to equal the corresponding conditional expected value, it is necessary for the economic agents to have perfect knowledge of the complete economic structure, except for the truly random disturbances. This means that they must know the values of all relevant parameters determining the underlying economic relationships, as well as the means of all exogenous variables, including exogenous government policy variables. While one can argue, as I shall, that given the availability of consistently estimated economic models, it may not be too unreasonable to assume that economic forecasters have unbiased estimates of relevant parameters, it is most unlikely that they will have such knowledge of exogenous variables, especially those under government control. Indeed, as I shall show below, in certain cases to be discussed, it is precisely in the government's interests to deliberately misinform the public as to its proposed policies if it wishes them to be effective. As a result of the overwhelming quantity of information it assumes, the use

*Professor of economics, Australian National University. An earlier version of this paper was presented to the Workshop in Macroeconomics at the University of Virginia; I wish to thank participants of the workshop for their helpful comments. The exposition of the paper has benefited from the suggestions of the managing editor and an anonymous referee.

of rational expectations has itself come under critical review recently (see, for example, Benjamin Friedman, Robert Gordon) especially in view of some of the rather dramatic implications it yields for short-run monetary and fiscal policy (see Thomas Sargent and Neil Wallace 1975, 1976).

While one might not want to take the extreme position implicit in the rational expectations hypothesis, nevertheless the basic idea it embodies, that expectations are generated from some underlying economic structure, is an extremely important one. With minor modification it can be adapted to provide useful insights into policy questions in a situation where less information than that required for rationality is assumed. In this paper we consider the implications of endogenizing expectations in such a manner for monetary and fiscal policy in a simple short-run macroeconomic model. We assume that forecasters have access to econometric models which provide them with unbiased estimates of the true structural relationships. The assumption of unbiasedness is a simplifying one which could be modified without difficulty. From these structural relationships of the economy they generate their predictions of the endogenous variables *conditional on their predictions of the exogenous variables.* In contrast to the rational expectations hypothesis, these exogenous predictions may be systematically wrong. Such will be the case if they are based on false information. If this is so, the endogenous forecasts, being based on the true structure, will also be systematically wrong. Given that the underlying exogenous information is incorrect, there is really nothing irrational in committing such systematic forecasting errors. Nonetheless, to avoid confusion with existing terminology I shall refer to our expectations as *structural* rather than rational. If the exogenous information is correct, then these structural forecasts will turn out to be rational in the Muthian sense. Moreover, as a simplification we shall avoid the introduction of random disturbances, so that if the exogenous forecasts are correct, the structural predictions will in fact hold exactly, rather than just on the average.[1]

My focus on the short run is intentional. The reason for doing so is that the issues raised by endogenizing forecasts in the above manner are most relevant over a relatively short time period. In the long run any autoregressive hypothesis used to predict endogenous variables will be correct and will therefore be rational. Likewise in the long run people should be able to predict sustained exogenous variables perfectly, so that their forecasts become rational as well. Indeed, as argued by Benjamin Friedman recently, the notion of rationality being based on perfect information (other than purely random errors) is essentially a long-run concept. It is in the short run that systematic forecasting errors may be committed and the government may have the scope to manipulate the public's expectations of its policies if it so wishes.

I. The Framework

Consider the following linear system of equations describing the reduced form solution at time t of a macroeconomic model

$$(1)\quad x(t) = Ax^*(t, t-1) + Bx^*(t+1, t-1) + Cz(t)$$

where

$x(t) = n \times 1$ vector of endogenous variables at time t

[1]Similar approaches to the one adopted in this paper can be found elsewhere in the literature. In the first place, our notion of structural expectations is similar to that of *consistent expectations* as defined by Alan Walters. Secondly, it is also close to that of *conditionally unbiased* predictions as used in econometric estimation; see for example Gordon Kaufman and F. W. McElroy. However, I prefer not to use these terms since they both suggest a complete stochastic specification, which I have not introduced. Thirdly, a more complete development of the underlying framework described in Section I below, is given by Robert Shiller who formulates his analysis within a fully stochastic context. His analysis of the "general linear rational expectations model" includes most of the conventional rational expectations models as special cases.

$x^*(t, t-1)$ = $n \times 1$ vector of predictions of the endogenous variables, formed *at* time $t-1$, *for* time t

$x^*(t+1, t-1)$ = $n \times 1$ vector of predictions of the endogenous variables, formed *at* time $t-1$, *for* time $t+1$

$z(t)$ = $m \times 1$ vector of exogenous variables at time t

A, B, C are $n \times n$, $n \times n$, $n \times m$, matrices of coefficients

Thus equation (1) can be interpreted as being a solution for a short-run linear (or linearized) macroeconomic model, in which the endogenous variables $x(t)$ are determined in part by their expectations, formed in the previous period for both the present and next period, and also by the exogenous variables $z(t)$, among which in general will be a subset of policy variables. As Section III below will show, recent macroeconomic models which stress the role of inflationary expectations are precisely of this form. Indeed they provided the rationale for adopting this formulation.

In addition to the arguments in (1), the solution may depend upon various lags of $x(t)$, $z(t)$, as well as expectations formed at various times in the more distant past, or for the more distant future. It is also possible, although not very common among conventional macroeconomic models, for the short-run solution to be dependent upon expectations of the exogenous variables. If so, the present framework can obviously be modified to incorporate them; only the complexity but not the substance of the analysis is changed. The matrices A, B summarize the short-run multipliers describing the impact effects of changes in expectations on the various endogenous variables for given values of the exogenous variables. The matrix C gives the short-run multipliers describing the impact of changes in the exogenous variables, given that these do not affect the expectations. Multipliers of this kind have been calculated recently in an explicitly expectational context by Thomas Sargent (1972), the author (1974), and the author and Andre Kaspura, and therefore fit precisely into the present framework. Other examples of such multipliers calculated in nonexpectational contexts abound in the literature and indeed form one of the standard methods of macroeconomic analysis.

Consider now expectations. We invoke the basic idea underlying the rational expectations hypothesis, namely, that forecasters base their expectations of endogenous variables on the predictions of the underlying economic model. For any arbitrary j-period forecast horizon we describe this by

$$(2) \quad x^*(t-1+j, t-1) = Ax^*(t-1+j, t-1) + Bx^*(t+j, t-1) + Cz^*(t-1+j, t-1) \quad \text{for } j = 1, 2, \ldots$$

where $x^*(t+j, t)$ = prediction of x formed at time t for time $t+j$

$z^*(t+j, t)$ = prediction of z formed at time t for time $t+j$

Thus the prediction formed at time $(t-1)$ for j time periods ahead is conditional on the corresponding forecasts of the exogenous variables, $z^*(t-1+j, t-1)$.

Applying equation (2) recursively yields the general solution

$$(3) \quad x^*(t-1+j, t-1) = \sum_{i=0}^{\infty} [(I-A)^{-1}B]^i (I-A)^{-1}C \; z^*(t-1+i+j, t-1) \quad \text{for } j = 1, 2, \ldots$$

In particular, setting $j = 1, 2$ (the forecasting horizons in (1)) yields

$$(4a) \quad x^*(t, t-1) = \sum_{i=0}^{\infty} [(I-A)^{-1}B]^i \cdot (I-A)^{-1}Cz^*(t+i, t-1)$$

$$\text{(4b)}\quad x^*(t+1,t-1) = \sum_{i=0}^{\infty} [(I-A)^{-1}B]^i \cdot (I-A)^{-1}Cz^*(t+1+i,t-1)$$

Thus the predicted value of x for time t formed at time $t-1$ will depend upon the predictions formed at time $t-1$ for the exogenous variable *for all future* periods. The same proposition holds for an arbitrary forecast interval j. Moreover, in order to ensure that these endogenous predictions remain finite we require that the eigenvalues of the matrix $(I-A)^{-1}B$ all lie within the unit circle. If $B=0$, so that only expectations formed for the present period are relevant, equation (3) simplifies to the contemporaneous relationship

$$\text{(3')}\quad x^*(t-1+j,t-1) = (I-A)^{-1}Cz^*(t-1+j,t-1) \qquad \text{for } j=1,2,\ldots$$

in which the predictions of x for j periods ahead depend upon the corresponding prediction for the exogenous variables.

It is an immediate consequence of (3) that if $z^*(t-1+i+j,\ t-1)$ is purely autoregressively determined, so that its value at time $t-1$ is predetermined, the conventional assumption of treating $x^*(t-1+j,t-1)$ as exogenous at time $t-1$ can be justified; it is simply a function of past $z(t-i)$ $(i \geq 1)$. This raises the question under what conditions can x^* be written as an autoregressive function, and, more specifically, when can it be described by an adaptive hypothesis? These issues, while of interest, are not central to the present discussion, and are therefore not pursued further here.[2]

[2]More formally, if $z_0(t)$ denotes the initial vector of exogenous variables, and $z_1(t)$ is the adjusted vector of exogenous variables, the change we are considering is defined by $dz(t) = z_1(t) - z_0(t)$ and measures a *shift* in the vector. In general this should not be confused with the alternative definition $dz(t) = z(t) - z(t-dt)$ which measures the change in the given vector z *between* two time periods. The one point when the two definitions do coincide is the period in which the change is first introduced. By focusing on the impact effects of any policy change, this is in fact the period we are considering. The same comments and interpretation apply to the expectations $dz^*(t-1)$ and to the other differentials defined below.

Inserting (4a) and (4b) into (1), with some manipulation we can write the solution for $x(t)$ in the form

$$\text{(5)}\quad x(t) = \sum_{i=0}^{\infty} [(I-A)^{-1}B]^i(I-A)^{-1}\cdot Cz^*(t+i,t-1) + C(z(t) - z^*(t,t-1))$$

thereby expressing $x(t)$ in terms of $z(t)$, together with the predictions for all future exogenous variables formed at time $t-1$.

Now consider equation (5) at time $(t-1+j)$. Comparing this to (3) we can calculate the forecasting error committed in predicting the endogenous variables j periods ahead. This is given by

$$\text{(6)}\quad x(t-1+j) - x^*(t-1+j,t-1) = \sum_{i=0}^{\infty} [(I-A)^{-1}B]^i(I-A)^{-1} \cdot C[z^*(t-1+i+j,t+j-2) - z^*(t-1+i+j,t-1)] + C(z(t-1+j) - z^*(t-1+j,t-1)) \qquad j=1,2,\ldots$$

and is seen to consist of two components. First there is the error in predicting the exogenous variables; secondly there is an error which may be induced through the updating of forecasts of the exogenous variables. In order for the predictions of the endogenous variables at a given time $t-1$ to be correct for any arbitrarily specified system (i.e., for arbitrary A, B, C), it is necessary and sufficient that

$$\text{(7a)}\quad z^*(t-1+j,t-1) = z(t-1+j)$$

$$\text{(7b)}\quad z^*(t-1+i+j,t+j-2) = z^*(t-1+i+j,t-1), \qquad \text{for } i=0,1,\ldots$$

That is, the j-period forecast of the exogenous variables made at time $t-1$ must be correct. Secondly, the forecasts made at time $t-1$ for time $t-1+j$ and each subsequent period beyond must be the same as that made for the same period at time $(t+j-2)$. This imposes certain restric-

tions on how forecasts of z may be revised, but does not preclude the possibility of some revision. Note that if $j = 1$, (7b) holds identically; the irrelevance of this condition can also be seen directly by comparing (3) and (4). Essentially, with a forecast horizon of only one period, the possibility of updating these forecasts obviously cannot arise.

If further, we require the j-period predictions of the endogenous variables to be correct for *all* forecasting dates t, it can easily be shown that (7a) must be strengthened to

$$(8)\quad z(t) = z^*(t, t-j) \qquad j = 1,2,\ldots$$

That is, the predictions of the exogenous variables for all periods t and for all planning horizons j must be correct.

As already noted, equation (5) expresses the current endogenous variables $x(t)$ in terms of

(i) the current *actual* exogenous variables $z(t)$;
(ii) the *predictions* of all exogenous variables, formed at time $t-1$, for time t and all subsequent periods.

Taking derivatives of (5), we obtain[3]

$$(9a)\quad \frac{\partial x(t)}{\partial z(t)} = C'$$

$$(9b)\quad \frac{\partial x(t)}{\partial z^*(t, t-1)} = [(I-A)^{-1}AC]'$$

$$(9c)\quad \frac{\partial x(t)}{\partial z^*(t+i, t-1)} = \{[(I-A)^{-1}B]^i(I-A)^{-1}C\}' \qquad i = 1,2,\ldots$$

where the prime denotes the vector transpose. The first effect is the impact of a change in the current exogenous variable $z(t)$ on $x(t)$ when no change is anticipated. This is precisely the expression obtained from (1), holding $x^*(t, t-1)$, $x^*(t-1, t-1)$ constant. The second effect describes the impact of a change in the predictions for the current period (formed in the previous period), with $z(t)$, $z^*(t+i, t-1)$ $(i \geq 1)$ held constant. It therefore represents the effect of an unsustained change in expectations, and operates through the expectational variable $x^*(t, t-1)$. An expected increase in an exogenous variable even if unrealized changes the expectations of the endogenous variables, which in turn will affect the actual outcome through equation (1). Equation (9c) describes the analogous effects arising from an unsustained change in expectations for any arbitrary future period, again as perceived at time $t-1$. These operate through the expectational variables $x^*(t, t-1)$, $x^*(t+1, t-1)$, and it will be observed that they decline at a geometric rate as they relate to the more distant future.

While equation (5) enables us to analyze the impact of a general time profile of predictions on the system, in practice, the expectations formed for the various future periods are most likely to be related. Thus as an important limiting case, we shall consider the situation where expectations formed at time $t-1$ are uniformly held for all future periods, so that

$$(10)\quad z^*(t-1+j, t-1) = z^*(t-1) \qquad j = 1,2,\ldots$$

with $z^*(t-1)$ denoting the (uniform) level of expectations held at time $t-1$. Substituting (10) into (3), the corresponding sustained endogenous expectations held at time $t-1$ are

$$(11)\quad x^*(t-1) \equiv x^*(t-1+j, t-1) = (I-A-B)^{-1}Cz^*(t-1) \qquad j = 1,2,\ldots$$

which are also therefore uniform throughout the future. Differentiating (11), the incremental effects of a change in exogenous expectations on the endogenous predictions are

$$(12)\quad \frac{\partial x^*(t-1)}{\partial z^*(t-1)} = \{(I-A-B)^{-1}C\}'$$

Similarly, it can be shown that (5) becomes

$$(13)\quad x(t) = (A+B)(I-A-B)^{-1} \cdot Cz^*(t-1) + Cz(t)$$

Thus the effect of a *sustained* change in ex-

[3] The derivative of the $(n \times 1)$ vector y with respect to the $(m \times 1)$ vector x, where y and x are related by $y = Mx$ and M is an $n \times m$ matrix is conventionally defined by $\partial y/\partial x = M'$.

ogenous expectations (i.e., sustained throughout all future periods) on $x(t)$ is given by

$$(14)\quad \frac{\partial x(t)}{\partial z^*(t-1)} = \{(A+B)(I-A-B)^{-1}C\}'$$

Indeed the impact of such a sustained anticipated change in the policy variables can be quite large relative to the impact of an actual policy change. This can be immediately seen in the case where A, B, and C are scalars (denoted by corresponding lower case letters) when

$$(15)\quad \frac{\partial x(t)/\partial z^*(t-1)}{\partial x(t)/\partial z(t)} = \frac{a+b}{1-(a+b)}$$

The ratio in (15) ranges between 0 and ∞ as $(a+b)$ increases from 0 to 1.

Henceforth, we shall focus our attention on sustained policy changes, the effects of which are expected to continue throughout the indefinite future. Let us therefore restrict ourselves to the simpler systems (11) and (13). In general, we shall hypothesize the following relationship

$$(16)\quad z^*(t-1) = \Gamma z(t)$$

where $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_n)$ and as in (10), $z^*(t-1)$ denotes the prediction of the exogenous variables made at time $t-1$, for the present time t. This forecast is expected to prevail throughout all future periods as well. Taking differentials of (16) yields

$$(17)\quad dz^*(t-1) = \Gamma dz(t)$$

which relates the change in the predictions of the exogenous variables (as measured from their initial levels) to the corresponding changes in the exogenous variables.[4] Given that $dz(t)$ is sustained, $(\Gamma - I)$ measures the accuracy with which changes in the exogenous variables are predicted. It is important to stress that equation (16) does *not* specify any mechanism generating the exogenous forecasts. All this equation describes is the extent to which actual exogenous changes are anticipated, irrespective of the forecasters' source of information.[5] If these expectations are formed autoregressively, then whether or not an increase in $z(t)$ is anticipated will be dependent upon the source of the change. For example, if $z(t)$ is itself autoregressively determined with a similar autoregressive structure to $z^*(t-1)$, then any increase in $z(t)$ generated by the autoregressive structure should be anticipated with a high degree of accuracy. On the other hand, if the increase in $z(t)$ is due to random disturbances, one would expect it to have very little immediate impact on expectations; the exogenous disturbance will need time to feed into the autoregressive process. However, the expectations $z^*(t-1)$ may be based on various other kinds of extraneous information, such as policy intentions announced by the government. The exact nature of the information can be left unspecified and is of no particular concern, at least in the present context.

Taking differentials of (13), and substituting (16) yields

$$(18)\quad dx(t) = [(A+B)(I-A-B)^{-1}C\Gamma + C]dz(t)$$

and using (12), can be written as

$$(19)\quad dx(t) = \left((A+B)\left(\frac{\partial x^*(t-1)}{\partial z^*(t-1)}\right)'\Gamma + C\right)dz(t)$$

Any actual exogenous change can therefore be seen to have two effects. First, there is the direct effect C as before. Second, to the extent that the change is anticipated, it will give rise to an induced expectational effect. With perfect forecasting $\Gamma = I$, so that (18) simplifies to[6]

$$(18')\quad dx(t) = (I-A-B)^{-1}Cdz(t)$$

[4] Friedman has considered this question in the context where forecasters have perfect knowledge of exogenous variables, but learn the true structural relationships through a least squares learning process. He shows that under certain conditions this can yield an adaptive type expectations hypothesis. The analogous question in the present context is discussed in an expanded version of this paper and is available from the author on request.

[5] A similar general approach to that undertaken here has been followed by Leif Johansen in his analysis of the use of forecasts in policy decisions.

[6] In the long run, of course, $\Gamma = I$ and expectations are rational.

II. Application to the Short-Run Macroeconomic Model

We now apply the framework discussed in Section I to a variant of the short-run macroeconomic model analyzed in my 1974 paper. This model consists of the following relationships:[7]

$$\text{(20a)}\quad Y(t) = C(Y(t)(1-u)) + I(r(t) - p^*(t+1,t-1)) + G(t) \qquad 0 < C' < 1,\ I' < 0$$

$$\text{(20b)}\quad L(Y(t), r(t), p^*(t+1,t-1)) = \frac{M}{P(t)} = \frac{M}{P(t-1)(1+p(t))} \qquad L_1 > 0,\ L_2 < 0,\ L_3 < 0$$

$$\text{(20c)}\quad p(t) = a_0 + a_1(Y(t) - \overline{Y}) + bp^*(t,t-1) \qquad a_1 > 0,\ 0 \leqq b \leqq 1$$

$$\text{(20d)}\quad \Delta M(t) + \Delta B(t) = [G(t) - uY(t)]P(t) = [G(t) - uY(t)]P(t-1)(1+p(t))$$

where
$Y(t)$ = real national income in period t
$C(t)$ = real consumption plans in period t
$I(t)$ = real investment plans in period t
$G(t)$ = real government expenditure in period t
u = rate of taxation (assumed to be proportional)
$r(t)$ = nominal interest rate at time t
$P(t)$ = price level at time t
$p(t) = [P(t) - P(t-1)]/P(t-1)$ = actual rate of inflation in period t
$p^*(t+1,t-1)$ = anticipated rate of inflation during period $t+1$, as expected at time $t-1$
$p^*(t,t-1)$ = anticipated rate of inflation during period t, as expected at time $t-1$
L = demand for real money balances in period t
$M(t)$ = nominal supply of money at time t
$B(t)$ = nominal stock of government bonds at time t (taken to be variable-interest rate bonds)
$\overline{Y}$ = full-employment level of real output

This is basically just a conventional *IS-LM* model, augmented by the inclusion of a Phillips curve embodying the "expectations hypothesis." Underlying it is the familiar notion of Hicksian- or Patinkin-like market periods, in which expenditure and asset plans for time t are made at time $t-1$.[8] Equation (20a) describes product market equilibrium, in which consumption depends upon disposable income and investment upon the real rate of interest. Given that investment plans for time t (i.e., to be put into effect over the period $(t, t+1)$) are made at time $t-1$, the relevant expected rate of inflation for the investment decision is $p^*(t+1, t-1)$, that expected at time $t-1$ for the next period $(t, t+1)$. Money market equilibrium is described by (20b), with the demand for money depending positively upon income and negatively upon both the nominal interest rate and the future expected rate of inflation.[9] A money demand function of the form in (20b) can be readily obtained if one follows James Tobin in assuming that the demand for real money balances depends positively on the real return on holding money (minus the rate of inflation) and negatively on the real return on holding bonds.[10] Note also for precisely the same reasons as investment, it is the expected rate

[7]Throughout the remainder of this paper a prime is used to denote the derivative of a function of one variable. Partial derivatives are indicated by numerical subscripts.

[8]To avoid confusion, time intervals must be defined carefully. All transactions are assumed to take place at the discrete points of time $t = 0, 1, 2$, etc. The time subscript on stocks therefore refers to the quantity at time t. For flows, the time subscript refers to period t, i.e., the unit time interval $(t-1, t)$.

[9]For simplicity we abstract from a fractional reserve banking system, so that $M(t)$ refers to the monetary base.

[10]For example, a Tobin-type money demand function would be $l[Y(t), r(t) - p^*(t+1, t-1), -p^*(t+1, t-1)]$ where $l_1 > 0$, $l_2 < 0$, $l_3 > 0$. Rewriting this as the function L in (20b), this corresponds to $L_1 = l_1$, $L_2 = l_2$, $L_3 = -(l_2 + l_3)$. Our assumption $L_3 < 0$, is therefore equivalent to $l_3 > -l_2$.

of inflation over the market period $(t, t+1)$ which is relevant. The actual rate of change of prices is postulated in (20c) to depend upon excess demand in the product market, as well as upon the expected rate of inflation which at time $t-1$ was expected to prevail for the *present* period $(t-1, t)$. The coefficient b measures the extent to which inflationary expectations are directly reflected in current price changes; the case $b = 1$ gives rise to the "natural rate of unemployment." The final equation describes the government budget constraint in nominal terms. The nominal deficit is financed either by printing money or by issuing bonds, which for simplicity are assumed to have variable interest rates.

This is an extremely simple one-period model, abstracting from such things as wealth effects in consumption and money demand, as well as interest payment on government debt, which form part of disposable income and contribute to the government deficit. The reason for abstracting from these factors is twofold. First, they really become important only in a longer run analysis. Most empirical studies suggest that wealth effects do not operate immediately. Moreover, I have shown elsewhere (1975) that in order to be consistent with the aggregate budget constraint of the economy, they *must* be introduced with a one-period lag. Hence they do not affect our first-period results, although they will begin to come into effect in the second period. Second, when they do become operative they introduce indeterminacies into the results, which tend to obscure the relevant issues which are our main concern.

The model is also similar in many respects to that used by Sargent (1973) and by Sargent and Wallace (1975). It is less general than theirs in the sense that they consider a fully stochastic system; it is more general in that it does not restrict the price expectations coefficient b to be unity and allows for biased forecasts of government policy. However, the two models are quite consistent and the system described by (20) yields deterministic analogues to the Sargent and Sargent-Wallace propositions when these more stringent conditions are imposed.

It is clear from (20d) that of the three government policy variables, $G(t)$, $\Delta M(t)$, $\Delta B(t)$, only two can be chosen independently, with the third being endogenously determined from this constraint. It is convenient for us to take the two independent variables to be $G(t)$ and $\Delta M(t)$, allowing $\Delta B(t)$ to be the residually determined variable. In this case, in the absence of wealth effects and interest payments, $\Delta B(t)$ has no feedback effects on the other endogenous variables $Y(t), r(t), p(t)$, so that there is no need to consider (20d) explicitly. We may therefore solve the system recursively in the sense that (20a)-(20c) together determine $Y(t), r(t), p(t)$ in terms of $p^*(t+1, t-1)$, $p^*(t, t-1)$, $M(t)$, $G(t)$ (and other exogenous parameters) with $\Delta B(t)$ then determined from (20d). Note also that since the system involves expectations for time t, as well as for time $(t+1)$, it is a non-linear form of the system postulated in (1).

With appropriate rearrangement of (20a)-(20c), the differential of this subset of equations can be written as equation (21).

$$(21)\quad \begin{pmatrix} 1 - C'(1-u) & -I' & 0 \\ -L_1 & -L_2 & -\dfrac{M(t)}{P(t)(1+p(t))} \\ -a_1 & 0 & 1 \end{pmatrix} \begin{pmatrix} dY(t) \\ dr(t) \\ dp(t) \end{pmatrix} = \begin{pmatrix} -I'dp^*(t+1, t-1) + dG(t) \\ L_3 dp^*(t+1, t-1) - dM(t)/P(t) \\ bdp^*(t, t-1) \end{pmatrix}$$

$$(23)\quad \begin{bmatrix} dY(t) \\ dr(t) \\ dp(t) \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dfrac{bI'M}{P(1+p)J} \\ 0 & 0 & \dfrac{[1-C'(1-u)]Mb}{P(1+p)J} \\ 0 & 0 & \dfrac{[-L_2[1-C'(1-u)]-I'L_1]b}{J} \end{bmatrix} \begin{bmatrix} dY^*(t,t-1) \\ dr^*(t,t-1) \\ dp^*(t,t-1) \end{bmatrix},$$

$$+ \begin{bmatrix} 0 & 0 & \dfrac{I'(L_2+L_3)}{J} \\ 0 & 0 & \dfrac{L_3[1-C'(1-u)]-a_1I'M/P(1+p)-L_1I'}{J} \\ 0 & 0 & \dfrac{a_1I'(L_2+L_3)}{J} \end{bmatrix} \begin{bmatrix} dY^*(t+1,t-1) \\ dr^*(t+1,t-1) \\ dp^*(t+1,t-1) \end{bmatrix}$$

$$+ \begin{bmatrix} \dfrac{-L_2}{J} & \dfrac{-I'}{PJ} \\ \dfrac{L_1+a_1M/P(1+p)}{J} & \dfrac{-[1-C'(1-u)]}{PJ} \\ \dfrac{-a_1L_2}{J} & \dfrac{-a_1I'}{PJ} \end{bmatrix} \begin{pmatrix} dG(t) \\ dM(t) \end{pmatrix}$$

The matrix on the left-hand side of (21), F say, which is just the matrix of partial derivatives of this subsystem, can easily be shown to be a P matrix. Thus invoking the David Gale-Hukukane Nikaido univalence theorem, it follows that this system can be solved uniquely for all endogenous variables, for any arbitrary set of values for the exogenous variables and parameters.[11] We can therefore write

$$(22a)\quad p(t) = f[p^*(t+1,t-1),\; p^*(t,t-1), M(t), G(t)]$$

$$(22b)\quad Y(t) = g[p^*(t+1,t-1),\; p^*(t,t-1), M(t), G(t)]$$

$$(22c)\quad r(t) = h[p^*(t+1,t-1),\; p^*(t,t-1), M(t), G(t)]$$

Premultiplying (21) by F^{-1}, the system can be solved for the changes in the endogenous variables in the form shown in equation (23) where

$$J = -L_2[1-C'(1-u)] - L_1I' - a_1I'M/P(1+p) > 0$$

Equation (23) is precisely of the form of (1) (expressed in changes rather than levels), so that the three matrices appearing on the right-hand side can be identified with A, B, C, respectively.[12] Their corresponding elements accordingly can be interpreted as in equation (1) above. In particular, C summarizes the effects of change in $G(t)$, $M(t)$ on the assumption that expectations remain unchanged. I shall henceforth refer to these as the *direct* effects, the signs of which are

[11] A square matrix is defined to be a P matrix if all its principal minors are positive. If this condition is met, the Gale-Nikaido global univalence theorem ensures that the system can be solved uniquely everywhere for the endogenous variables; i.e., *globally*. It is a much stronger condition than the nonvanishing of the Jacobian, which ensures only *local* uniqueness.

[12] Since equation (23) has been obtained by linearizing the system (20), we can consider only local stability. The partial derivatives and expressions appearing in (23) are therefore being evaluated at their equilibrium values.

straightforward and do not require any elaboration.[13]

In order for (23) with endogenously determined expectations to be stable, we have seen that the matrix $(I - A)^{-1}B$ must have eigenvalues lying in the unit circle. For the matrices A, B in (23), it can be seen that two of the eigenvalues of $(I - A)^{-1}B$ are zero; the remaining eigenvalue can be shown to be positive and the condition for it to lie within the unit circle is

$$(24) \qquad a_1 I' \frac{M}{P} \Omega < (1 - b)J$$

where $\Omega = \frac{e_r}{r} + \frac{e_{p^*}}{p^*} + \frac{b}{1 + p}$

and e_r = interest elasticity of the demand for real money balances $(\equiv (L_2 r)/(M/P))$

e_{p^*} = elasticity of the demand for real money balances with respect to inflationary expectations $(\equiv (L_3 p^*(t + 1, t - 1))/(M/P))$

The expression Ω played an important role in the comparative statics discussed by the author (1974), which treated expectations as an exogenous parameter. There it was argued that on the basis of plausible parameter values one would expect $\Omega < 0$.[14]

In this case an increase in (exogenous) inflationary expectations would increase real income, thereby increasing the actual rate of inflation in excess of b (the direct effect from the Phillips curve), and would lower the real rate of interest. In the event that $\Omega > 0$, the directions of these effects would be reversed.

With structurally determined expectations, however, we see that stability considerations in fact impose a restriction on Ω. If $b < 1$, then $\Omega < 0$ is quite compatible with stability; indeed the smaller b, the more likely it is that both the stability condition (24) and $\Omega < 0$ hold. If $b = 1$, on the other hand, $\Omega > 0$ becomes both necessary and sufficient for stability. Furthermore for $b = 1$, the likelihood of $\Omega > 0$ being met also increases. While for the parameters quoted in footnote 14, $b = 1$ implies $\Omega = -0.27$, and hence instability, it should be noted that these estimates were obtained over periods of relatively low inflation and low nominal interest rates, when estimates of b also tended to be low. Recent empirical evidence would suggest that values of b close to unity occur in periods of high inflation. Thus for example, taking $r = .08$, $p^* = 0.08$ as more representative of recent experience, and doubling e_{p^*} to say -0.02, we find that $\Omega = 0.30$, ensuring stability. In any event, irrespective of this empirical evidence, we require (24) to hold and henceforth shall include this condition among our restrictions.

As indicated in Section I, we shall be concerned with analyzing the effects of sustained policy changes. An important component of these is contained in the matrix $(I - A - B)^{-1}C$, which summarizes the incremental effects of changes in the predictions of exogenous variables on the structural expectations; see (12). With A, B, C as defined in (23), these induced expectations effects are given by equation (25) where the elements a_{ij}, b_{ij}, c_{ij} are all given in (23). From (23) and the stability condition (24), it can be seen that

$$(26) \qquad 1 > a_{33} + b_{33} > 0$$

ensuring that the denominator of the various elements in (25) are all positive.

[13]See the author (1974), where they are discussed in more detail. To summarize these results, we see that an increase in $M(t)$ will raise $Y(t)$, $p(t)$, and lower $r(t)$; an increase in $G(t)$ will raise all three endogenous variables.

[14]The argument that $\Omega < 0$ was based on the following considerations. Typical estimates of the short-run interest elasticity of the demand for money are about -0.03. There is much less available evidence on e_{p^*} and we take the estimates of -0.01 obtained by Lawrence Smith and John Winder. Assuming $r = 0.04$, $p^* = 0.02$, which are the averages over the sample periods upon which these estimates are based, we obtain $e_r/r = -0.75$, $e_{p^*}/p^* = -0.50$ so that $\Omega < 0$ for all feasible values of b in range $0 < b < 1$. More recent estimates of demand for money functions involving the nominal interest rate and the anticipated rate of inflation by Stephen Goldfeld tend to confirm these magnitudes. However, we definitely cannot rule out the possibility that $\Omega > 0$. Indeed I argue below that these estimates were obtained over periods of relative price stability, when r and p^* are low relative to their recent values. Increasing the values of r, p^* and taking $b = 1$, the possibility of $\Omega > 0$ now becomes quite plausible.

$$(25)\quad \begin{pmatrix} dY^*(t-1) \\ dr^*(t-1) \\ dp^*(t-1) \end{pmatrix} = \begin{pmatrix} c_{11} + \dfrac{(a_{13}+b_{13})c_{31}}{1-(a_{33}+b_{33})} & c_{12} + \dfrac{(a_{13}+b_{13})c_{32}}{1-(a_{33}+b_{33})} \\ c_{21} + \dfrac{(a_{23}+b_{23})c_{31}}{1-(a_{33}+b_{33})} & c_{22} + \dfrac{(a_{23}+b_{23})c_{32}}{1-(a_{33}+b_{33})} \\ \dfrac{c_{31}}{1-(a_{33}+b_{33})} & \dfrac{c_{32}}{1-(a_{33}+b_{33})} \end{pmatrix} \begin{pmatrix} dG^*(t-1) \\ dM^*(t-1) \end{pmatrix}$$

The elements of the matrix in (25) summarize the effects of changes in the *anticipated* money supply and government expenditure on the *anticipated* rate of inflation, level of income, and nominal interest rate. It is of some interest to compare these anticipated effects with the corresponding *actual* effects (what we have been calling the direct effects) summarized by the elements of C.

First, it is an immediate consequence of (25) that the effects of an anticipated expansionary government policy (either an increase in $G^*(t-1)$ or $M^*(t-1)$) will raise the anticipated rate of inflation. Moreover, its effect is stronger than the corresponding direct effects of actual policy changes on the actual rate of inflation. The reason is that the latter, given by c_{31} and c_{32} in (23), do not embody any induced expectational impact; these are explicitly incorporated in (25) through the term $1/[1-(a_{33}+b_{33})]$.

Secondly, it can readily be shown that if $b < 1$ and $\Omega < 0$, the effect of an anticipated increase in the quantity of money or government expenditure on the expected level of income will also be stronger than the corresponding direct effect measured by c_{11}, c_{12}, respectively. On the other hand, if $b = 1$ so that $\Omega > 0$ (as required by the stability condition (24)), anticipated government policy will have *no* effect on expected income. The reason is that with $b = 1$, $Y^*(t-1)$ is expected to remain at its natural rate and is therefore independent of any government policy.[15]

Thirdly, the effect of an increase in $M^*(t-1)$ on the expected nominal interest rate is indeterminate. While the direct effect, $c_{22} < 0$, an increase in $M^*(t-1)$ will also raise the anticipated rate of inflation, the effect of which is almost certainly to raise $r^*(t-1)$, thereby making the net effect ambiguous.[16] With $b = 1$, however, and $Y^*(t-1)$ independent of $M^*(t-1)$, it follows from the (expected) product market equilibrium condition that the expected real rate of interest must also be independent of $M^*(t-1)$. Thus since an increase in $M^*(t-1)$ will raise the expected rate of inflation, the expected nominal interest rate must rise as well. By contrast an expected increase in $G^*(t-1)$ holding $M^*(t-1)$ constant (i.e., bond financed) will raise the expected nominal interest rate unambiguously.

Finally, substituting $(I-A-B)^{-1}C$ into (13), we can express the changes in the endogenous variables in terms of the simultaneous changes in the exogenous policy variables and their expectations. These are given by equation (27). The first matrix of (27) summarizes the consequences of the induced expectational effects just discussed; the second matrix C simply gives the direct effects, as before. Equation (27) forms the basis of the analysis of policy changes, to which we now turn.

[15]This can be seen by using the definition in (23) to evaluate the elements

$$c_{1j} + \frac{(a_{13}+b_{13})c_{3j}}{1-(a_{33}+b_{33})} \qquad j = 1, 2 \text{ when } b = 1$$

[16]There is a slight ambiguity in this statement arising from the fact that my assumptions do not suffice to ensure $a_{23}+b_{23} > 0$. To the extent $a_{23}+b_{23}$ is ambiguous the impact of an increase in inflationary expectations (treated as an exogenous parameter) on the expected rate of interest is indeterminate. However, for plausible parameter values one would expect with some confidence that $a_{23}+b_{23} > 0$, creating the overall ambiguity discussed in the text.

$$(27)\quad \begin{pmatrix} dY(t) \\ dr(t) \\ dp(t) \end{pmatrix} = \begin{bmatrix} \frac{(a_{13}+b_{13})c_{31}}{1-(a_{33}+b_{33})} & \frac{(a_{13}+b_{13})c_{32}}{1-(a_{33}+b_{33})} \\ \frac{(a_{23}+b_{23})c_{31}}{1-(a_{33}+b_{33})} & \frac{(a_{23}+b_{23})c_{32}}{1-(a_{33}+b_{33})} \\ \frac{(a_{33}+b_{33})c_{31}}{1-(a_{33}+b_{33})} & \frac{(a_{33}+b_{33})c_{32}}{1-(a_{33}+b_{33})} \end{bmatrix} \begin{pmatrix} dG^*(t-1) \\ dM^*(t-1) \end{pmatrix} + \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} \begin{pmatrix} dG(t) \\ dM(t) \end{pmatrix}$$

III. Monetary and Fiscal Policy

We now consider the short-run effects of changes in monetary and fiscal policy on the endogenous variables of the model.

A. *Open Market Operation*

Assume for simplicity that the government deficit at time $t-1$, $D(t-1) = [G(t-1) - uY(t-1)] = 0$. The budget constraint at time t can then be written in differential form as[17]

$$(20d')\quad dM(t) + dB(t) = [dG(t) - udY(t)]P(t)$$

Setting $dG(t) = 0$, an open market operation is thus described by

$$(28)\quad dB(t) = -dM(t) - udY(t)P(t)$$

Treating $dM(t)$ as the independent policy variable, this equation describes the required transaction of bonds given the initially balanced budget and the induced tax receipts, which will follow from the resulting change in income.

We shall assume that $dM^*(t-1)$ and $dM(t)$ are related by

$$(29)\quad dM^*(t-1) = \delta dM(t)$$

so that $(\delta - 1)$ measures the accuracy with which the change in money supply is predicted. It is overpredicted, perfectly predicted, or underpredicted according as $\delta \gtrless 1$. If $\delta < 0$, the direction, as well as the magnitude of the change is incorrectly predicted. Inserting (29) into (27), the short-run impacts of changes in the money supply can be obtained (where for simplicity time subscripts are dropped in the resulting expressions):

$$(30a)\quad \frac{dp(t)}{dM(t)} = c_{32}\left(\frac{(a_{33}+b_{33})\delta}{1-(a_{33}+b_{33})} + 1\right) = c_{32}\left(\frac{\delta(bJ + I'a_1M\Omega/P)}{(1-b)J - I'a_1M\Omega/P} + 1\right)$$

$$(30b)\quad \frac{dY(t)}{dM(t)} = \frac{c_{32}(a_{13}+b_{13})\delta}{1-(a_{33}+b_{33})} + c_{12} = c_{12}\left(\frac{\delta a_1 I'\Omega M/P}{(1-b)J - I'a_1M\Omega/P} + 1\right)$$

$$(30c)\quad \frac{dr(t)}{dM(t)} = \frac{c_{32}(a_{23}+b_{23})\delta}{1-(a_{33}+b_{33})} + c_{22}$$

Let us first consider the effect on the short-run rate of inflation. This consists in part of the direct effect c_{32}. In addition, to the extent that the change is anticipated, there is an induced effect which operates through the endogenous inflationary expectations. Given the fact that $1 > a_{33} + b_{33} > 0$, this is always positive as long as the direction of the change is correctly anticipated ($\delta > 0$). That is, the anticipated increase in the quantity of money will raise the anticipated rate of inflation in accordance with (25), and this in turn will increase further the actual rate of inflation. Note that as commented in Section I, the impact of this expectations-induced effect may be quite

[17]In interpreting the government budget constraint written in differential form, the comments made in fn. 2 should be borne in mind.

large relative to the direct effect, especially if the policy change is predicted with accuracy. Taking $\delta = 1$, $\Omega < 0$, and $b = 0.8$ (a reasonable estimate in the light of empirical evidence), then

$$\frac{\partial p(t)/\partial M^*(t-1)}{\partial p(t)/\partial M(t)} > \frac{b}{1-b} = 4$$

Thus the induced effect is more than *four* times that of the direct effect. This does have some policy implications. It means that if the government wishes a contractionary monetary policy to have maximum impact on reducing the rate of inflation, it should inform the people that such a restrictive policy will be introduced. That way it will be able to generate significant expectational effects which will contribute substantially to the desired impact of the policy change.[18]

The effect on output is given in (30b). Suppose $\Omega < 0$, $\delta > 0$; then both the direct effect and the induced-expectations effect of an anticipated increase in $M(t)$ will be expansionary. An increase in the anticipated supply of money will raise the expected rate of inflation, and thereby investment and output. And the more people anticipate the expansionary policy, the more expansionary it will be.

Taking $b = 1$, however,

$$\frac{dY(t)}{dM(t)} = c_{12}(1-\delta)$$

In this case, the expectations effect will be offsetting. The reason is that in order to ensure stability, it is now necessary for $\Omega > 0$, the consequence of which is for the induced increase in the expected rate of inflation to have a contractionary effect. If, further, $\delta = 1$, we see that a change in the money supply will have no effect on output; the induced expectations effect will exactly offset the conventional direct effect. With $\delta = 1$, all endogenous variables will be predicted perfectly, so that expectations are rational in the Muthian sense. Thus from (20c) and (25) we see that $Y(t) = Y^*(t-1) = \bar{Y}$ and is fixed. And this is precisely the result obtained by Sargent (1973) and Sargent and Wallace (1975) referred to earlier. If expectations are rational and the natural rate hypothesis holds, monetary policy is ineffective in controlling output *even in the short run*.[19] Put another way, if $b = 1$, the only way monetary policy can influence output in the conventional way is if the government can succeed in making people underpredict the policy change (i.e., $\delta < 1$). On the other hand, if $\delta > 1$, its effects will be perverse.

For $\Omega < 0$, $\delta > 0$, the effect of a change in $M(t)$ on the nominal interest rate is indeterminate. The reason is that while the direct effect of an increase in $M(t)$ will be to lower it, the increase in inflationary expectations so generated will (almost certainly) tend to raise it.[20] In this case, if the monetary authorities wish to lower the interest rate by means of an expansion in the money supply, then in order for this policy to be effective, it should deliberately try to mislead the public into believing that a contractionary policy is about to take place. That way, the offsetting expectational element will operate in the desired direction. In the natural rate case $b = 1$, and with perfect predictions $\delta = 1$, both effects operate in the same direction making $dr(t)/dM(t) > 0$ unambiguously.

It is also of interest to consider the effect on the real rate of interest $r_b(t)$ say $= r(t) - p^*(t-1)$. This is simply

$$(31)\quad \frac{dr_b(t)}{dM(t)} = c_{22}\left(1 - \frac{M\Omega c_{32}\delta}{1-(a_{33}+b_{33})}\right)$$

For $\Omega < 0$, $\delta > 0$ this will be unambiguously negative. But if $b = \delta = 1$ so that conditions for both the natural rate of un-

[18]It is worth noting that the main reason for the large induced expectational effect is due to the *sustained* nature of the change in policy.

[19]This is the deterministic analogue to Sargent's first proposition that "... a natural rate of output exists in the sense that the deviation of output from its normal level is statistically independent of the systematic parts of monetary and fiscal policies; ..." (1973, p. 442).

[20]In this connection the indeterminacy noted in fn. 16 above should be borne in mind.

employment and rational expectations prevail, then it can readily be shown from (31) the real rate of interest is independent of the money supply, consistent with a second proposition obtained by Sargent (1973).[21]

B. *Bond-Financed Increase in Government Expenditure*

This case is considered by setting $dM(t) = 0$ in (20d′) so that

$$dB(t) = [dG(t) - udY(t)]P(t) \tag{32}$$

The results it yields are virtually identical to those we have been discussing for an open market operation. Assuming that $dG^*(t-1)$ and $dG(t)$ are related by

$$dG^*(t-1) = \mu dG(t)$$

where $(\mu - 1)$ measures the accuracy with which $dG(t)$ is predicted, we obtain

$$\text{(33a)}\quad \frac{dp(t)}{dG(t)} = c_{31}\left(\frac{(a_{33}+b_{33})\mu}{1-(a_{33}+b_{33})} + 1\right) = c_{31}\left(\frac{\mu(bJ + I'a_1M\Omega/P)}{(1-b)J - I'a_1M\Omega/P} + 1\right)$$

$$\text{(33b)}\quad \frac{dY(t)}{dG(t)} = \frac{c_{31}(a_{13}+b_{13})\mu}{1-(a_{33}+b_{33})} + c_{11} = c_{11}\left(\frac{\mu a_1 I'\Omega M/P}{(1-b)J - I'a_1M\Omega/P} + 1\right)$$

$$\text{(33c)}\quad \frac{dr(t)}{dG(t)} = \frac{c_{31}(a_{23}+b_{23})\mu}{1-(a_{33}+b_{33})} + c_{21}$$

The effects on the current rate of inflation and level of output are identical to (30a) and (30b), respectively, except for a constant of proportionality $c_{31}/c_{32} = c_{11}/c_{12}$. Therefore all the comments made above on the effects of monetary policy on $p(t)$, $Y(t)$ apply directly to the effects of increases in government expenditure. The effects on the nominal interest rate do not carry over directly, however. Provided $\mu > 0$, both the direct effect of an increase in $G(t)$ and the induced inflationary effect operate in the same direction, making $dr(t)/dG(t) > 0$ unambiguously. For $\Omega > 0$, and $\mu > 0$, the induced expectational effect on the real rate of interest is offsetting, making the net effect ambiguous. This indeterminacy disappears when $b = 1$, in which case the real rate of interest will rise. The economic reasoning for this result can be seen most clearly when $\mu = 1$ as well. Under these circumstances an increase in government expenditure will have no effect on income or on consumption. Therefore, the only way that the increase in government expenditure can be realized is by "crowding out" an equal quantity of private investment; that is, for the real interest rate to rise. Again this last result is consistent with the corresponding conclusion obtained previously by Sargent.

C. *Money-Financed Increase in Government Expenditure*

These effects are essentially a combination of the policies considered under the preceding sections A and B. The actual change in the money supply is

$$dM(t) = [dG(t) - udY(t)]P(t) \tag{34}$$

with a corresponding equation-linking expectations. Inserting these relationships into (30) and (25) enables us to solve for the corresponding impacts of $dG(t)$ on the endogenous variables. Because of the fact that now $dM(t)$ (rather than just $dB(t)$) depends upon the induced tax receipts in (34), the calculations are somewhat more complex than before, although the results are precisely as one would expect. The effects on the rate of inflation and level of income continue to apply as before. The effects on both the nominal and real interest rates are indeterminate unless $b = 1$, and the policy is perfectly anticipated. The reason is that while the effect of the increase in $G(t)$ is to raise $r(t)$, the increase in $M(t)$ has an ambiguous impact. With $b = 1$ and Muthian rational expectations, both the monetary and fiscal effects operate in the same direction (or at least are not offsetting) implying that a money-financed increase in government expenditure will raise both the nominal and real interest rate.

[21]The proposition as stated by Sargent is "... the real rate of interest is independent of the systematic part of the money supply; ..." (1973, p. 443).

IV. Summary

This paper has considered a short-run macroeconomic model in which the expectations of endogenous variables are determined by the structure of the model, conditional on the given exogenous predictions of the exogenous variables. It is shown how any change in government policy gives rise to two effects. First there is the conventional direct effect, which ignores any impact the policy may have on these endogenous expectations. Second, to the extent that the policy change is anticipated, it gives rise to an induced expectations effect, which operates through its impact on endogenous expectations. This analysis is applied to a simple macroeconomic model in which the only endogenous expectations variable is the expected rate of inflation. The main results are as follows.

We show that provided the process generating the structural expectations is stable, and as long as the direction of the policy change is correctly predicted, the induced expectations effect of either an increase in the money supply or an increase in government expenditure (the latter with either a bond-financed or money-financed government deficit) will reinforce the direct positive effect on the current rate of inflation. In this case if the government wishes to use say a restrictive monetary policy to reduce the rate of inflation, it can increase the effectiveness of the policy by announcing it in advance. This will enable it to get incorporated into the public's expectations, thereby increasing the induced expectations effect.

If $b < 1$ and $\Omega < 0$, conditions which empirical evidence suggests are likely to be met at least in periods of low inflation, the same proposition applied with respect to the effectiveness of monetary and fiscal policy on the level of income; both the direct and induced expectations effects operate in the same direction. On the other hand, if $b = 1$ so that the natural rate hypothesis holds, $\Omega > 0$ is necessary (and sufficient) for stability, and the induced-expectational effect is offsetting. In this case, in order to increase the effectiveness of its policies, the government should misinform the public as to its plans. Indeed if the government policy is predicted perfectly, it will be ineffective in controlling income, even in the short run. The only way it can be successfully used to change income is if the public underpredict the policy change; if they overpredict, its effects will be perverse.

The impacts of the direct and induced expectational effects on the interest rate are slightly more complicated. With $\Omega < 0$, and the direction of the policy change correctly anticipated, (i.e., $\delta > 0$), the two effects of an expansionary monetary policy on the nominal interest rate are offsetting, while their net effect on the real rate of interest is unambiguously negative. With $b = 1$ and with perfect predictions, the real rate of interest is independent of the nominal money supply, implying that an expansionary monetary policy will raise the nominal interest rate. On the other hand, for $\mu > 0$, the two effects of an increase in government expenditure with a bond-financed deficit will tend to raise the nominal interest unambiguously. For $\Omega < 0$, its effects on the real rate are ambiguous, although for $b = 1$, the real rate will rise as well. With a money-financed deficit, and $\Omega < 0$, the net effect of an increase in government expenditure on both the nominal and real rate of interest will be indeterminate; however, for $b = 1$ and perfect anticipation, both the nominal and real interest rates will rise.

REFERENCES

B. M. Friedman, "Rational Expectations are Really Adaptive After All," unpublished work. paper, Harvard Univ. 1975.

D. Gale and H. Nikaido, "The Jacobian Matrix and Global Univalence of Mappings," *Mathemat. Ann.*, 1965, *159*, 81–93.

S. M. Goldfeld, "The Demand for Money Revisited," *Brookings Papers*, Washington 1973, *3*, 577–646.

R. J. Gordon, "Recent Developments in the Theory of Inflation and Unemployment," *J. Monet. Econ.*, Apr. 1976, *2*, 185–220.

L. Johansen, "On the Optimal Use of Forecasts in Economic Policy Decisions," *J. Publ. Econ.*, Apr. 1972, *1*, 1–24.

G. M. Kaufman, "Conditional Prediction and Unbiasedness in Structural Equations," *Econometrica*, Jan. 1969, *37*, 44–49.

F. W. McElroy, "Unbiased Estimation of Conditional Expectations," *Int. Econ. Rev.*, Oct. 1971, *12*, 517–18.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

T. J. Sargent, "Anticipated Inflation and the Nominal Rate of Interest," *Quart. J. Econ.*, May 1972, *86*, 212–25.

———, "Rational Expectations, the Real Rate of Interest and the Natural Rate of Unemployment," *Bookings Papers*, Washington 1973, *2*, 429–72.

——— **and N. Wallace**, "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–57.

——— **and** ———, "Rational Expectations and the Theory of Economic Policy," *J. Monet. Econ.*, Apr. 1976, *2*, 169–84.

R. Shiller, "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review," presented at Conference on the Monetary Mechanism in Open Economies, Helsinki 1975.

L. B. Smith and J. W. L. Winder, "Price and Interest Expectations and the Demand for Money in Canada," *J. Finance*, June 1971, *26*, 671–82.

J. Tobin, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15–29.

S. J. Turnovsky, "On the Role of Inflationary Expectations in a Short-Run Macroeconomic Model," *Econ. J.*, June 1974, *84*, 317–37.

———, "Monetary Policy, Fiscal Policy and the Government Budget Constraint," *Australian Econ. Pap.*, Dec. 1975, *14*, 197–215.

——— **and A. Kaspura**, "An Analysis of Imported Inflation in a Short-Run Macroeconomic Model," *Can. J. Econ.*, Aug. 1974, *7*, 355–80.

A. A. Walters, "Consistent Expectations, Distributed Lags and the Quantity Theory," *Econ. J.*, June 1971, *81*, 273–81.

Public Services, Private Substitutes, and the Demand for Protection Against Crime

By CHARLES T. CLOTFELTER*

It is a common notion in the theory of public finance that privately provided goods may be substituted for publicly provided goods, such as locks for more police and sprinkler systems for more firemen. Yet empirical studies of the demand for public goods have in general ignored such substitutes. For example, Theodore Bergstrom and Robert Goodman correctly note that the prices of private goods affect individual demand for public goods, p. 283, but in their empirical analysis they assume that these prices and the unit prices of public goods are the same in all communities. The possibility of public-private substitution has received some theoretical attention. John Head and Carl Shoup touch on this problem in their normative analysis of the proper identification of public, private, and "ambiguous" goods, based on each good's optimal allocation. And in his text, Shoup, p. 68, mentions the simultaneous provision of burglar alarms and police within the context of his discussion of public services. In another application to crime prevention, Henry Tulkens and Alex Jacquemin present a theoretical analysis which allows for the simultaneous provision of public and private protection. It is primarily in the area of education, however, that the public-private choice has received much positive analysis (see Mark Pauly, Yoram Barzel, Sam Peltzman, and Joseph Stiglitz). The purpose of this paper is to analyze the role of private substitutes for public services within the more general context of public sector decision making, and then to apply this analysis in an empirical analysis of the demand for protection against crime. Section I illustrates the importance of the private substitutes available for some local public services and describes the various forms of private protection. Section II presents a general analysis of the choice between private and public modes of provision. Section III illustrates this analysis with an application to the case of private protection against crime and presents estimates of a *CES* form of the production of security using data for *U.S.* states in 1970. The implications of these findings for rising expenditures by local governments are also noted. Section IV summarizes the analysis.

*University of Maryland. I am grateful to Henry Aaron, Ann Bartel, Theodore Bergstrom, George Borts, Charles Brown, Harry Kelejian, A. Thomas King, Bruce Vavrichek, Randall Weiss, and members of the Maryland Public Economics Seminar for helpful comments and discussions; to Judith Radlinski for research assistance; and to the University of Maryland Computer Science Center for research support. After this paper was completed, I learned of a doctoral dissertation dealing with similar issues for the case of fire protection. The author is Charles Vehorn. The thesis was submitted to Ohio State University.

I. Background

Private substitutes for a number of public services have become increasingly important in this country. Examples of such private substitutes are familiar: private schools for public schools; locks, alarms, and protective behavior for public police; private swimming pools, yards, private clubs for public recreational facilities; personal copies for public library books. The importance of these private substitutes is in some cases quite great. As an indication: over 10 percent of elementary and secondary pupils in 1970 were enrolled in nonpublic schools;[1] state and local governments spent $7.5 billion on police and fire protection in 1971[2] while estimates of annual private expenditures for protection ranged from $3 billion to $15 billion for recent

[1]See U.S. Office of Education, p. 7.
[2]See Tax Foundation, p. 136.

years;[3] local governments in 1971 spent $2.1 billion on parks and recreation[4] while Americans spent an estimated $1.5 billion on club dues and memberships alone (not to mention $25.2 billion more for other recreation);[5] and private itemized contributions to the United Fund in 1972 totaled $0.6 billion,[6] though that figure was far surpassed by local public welfare expenditures of $7.7 billion.[7] While private substitutes may be of increasing importance in some areas, a glance at the history of local and national public spending reveals numerous functions passing from the private to the public domain. During the nineteenth century, for example, local governments began to take responsibility for education, police protection, sewage, and libraries, among other functions.[8]

This paper applies the analysis of the public-private choice of provision to the demand for protection against crime, a problem of growing economic importance. Between 1960 and 1970 the rate of reported violent crimes increased by 146 percent and the corresponding rate for property crime rose 160 percent.[9] Although the interpretation of such reported crime rates is a matter of debate, it is widely agreed that there has been a significant growth in crime. At the same time, households and firms have turned to private means of protecting themselves on an unprecedented scale. The means of protection have included devices such as alarm systems, safes, automatic telephone dialers, window bars, and bigger locks as well as services such as those provided by private protective and detective agencies. Over the period 1964–73 employment in such agencies (SIC 7393) increased 226 percent compared to an increase of 54 percent for public police. Both of these rates of growth exceeded the 34 percent increase in all employment covered by the U.S. Bureau of the Census, *County Business Patterns (CBP)*.[10] Other forms of self-protection have experienced growth as well. Because aggregate data are not available for all of these forms of private protection, the empirical analysis presented in this paper employs data for private protective and detective services only. These data are discussed in more detail below.

II. The Public-Private Choice of Provision

Rather than assume that households directly value units of publicly provided goods per se—a usual characteristic of studies of the demand for public goods—it seems more realistic to view households as valuing the ultimate "output" to which public goods and services contribute.[11] This output q is measured in terms of units valued by households, such as recreational opportunities, educational quality, or the level of protection against crime in the community.[12] A typical household values this output as well as other goods (z): $U_i = U_i(z_i, q_i)$. The output q_i is produced with the use of a privately provided (market) good y_i and a publicly provided (nonmar-

[3]See *U.S. News and World Report*; Reginald Stuart, pp. F1, F7; and James Kakalik and Sorrel Wildhorn.

[4]See Tax Foundation, p. 233.

[5]Estimated by multiplying fraction of consumption devoted to recreation in the 1960–61 *BLS Survey* by total personal consumption in 1971 ($667.2 billion); U.S. Council of Economic Advisors, p. 263.

[6]See Allen Lerman, Table 12.

[7]See Tax Foundation, p. 233.

[8]For a description of the growing role of government in *U.S.* cities, see Blake McKelvey. For a case study of public decision making in the adoption of public education, see Michael Katz.

[9]See U.S. Bureau of the Census, *Statistical Abstract*, p. 143.

[10]The 1964 and 1973 levels of employment are, respectively: 62,170 and 202,561, industry 7393; 334,400 and 515,811, state and local police; 45,641,167 and 61,275,142, total employment covered in *CBP*. Figures on public police are in full-time equivalent units, but those from the *CBP* are not, so the comparison in rates of growth would be biased if the proportion of part-time employees has been changing in private protective agencies. However, there is no evidence of such a trend. Not included in total employment covered in *CBP* are workers in government, agriculture, domestic service, and railroad employment. See U.S. Bureau of the Census (1971a, b).

[11]This assumption is in direct contrast to William Baumol's characterization of production in the "nonprogressive" (including most of the public) sector in which "for all practical purposes the labor is itself the end product" (p. 416).

[12]For a similar approach, see David F. Bradford et al.

ket) good x_i: $q_i = g(x_i, y_i)$. Because they are used as inputs in the production of the ultimate output q_i, both x_i and y_i may usefully be thought of as intermediate goods.[13] If there is no discrimination, the amount of the publicly provided good available to all households is related to the level of community production X by the conventional expression: $x = XN^{-\alpha}$, where N is the population of the community and α is a measure of the "publicness" of the good which has the value of zero for a pure public good and one for a private good.[14]

For the community, the aggregate perceived amount of the publicly provided input is $Nx = XN^{1-\alpha}$. Assuming that household production relationships can be aggregated to the community level,[15] the production of the output Q for the community may be expressed as

$$(1) \quad Q = F(XN^{1-\alpha}, Y) \qquad \partial Q/\partial X \geq 0 \quad \partial Q/\partial Y \geq 0$$

where $XN^{1-\alpha}$ and Y are the total amounts of the publicly and privately provided goods, respectively.[16] Like all production functions, equation (1) gives the maximum output for a given set of inputs. This implies that an intermediate good such as public police—whose possible functions vary from investigating murders to dispensing parking tickets—are used so as to maximize the community's desired output. To the extent that resources are not efficiently allocated in practice, estimates of this model from real world data will be biased.

For simplicity, it is assumed that the share of community income spent on Q is a constant, independent of the costs of X and Y. This would be true, for example, if community demand could be described by a Cobb-Douglas utility function homogeneous of degree one. The community therefore faces a budget constraint of the form

$$(2) \qquad B(X, Y) = B^0$$

where B^0 is a constant level of expenditure. If units of the public and private inputs can be purchased at constant unit cost, this budget constraint is a straight line, but if there are economies of scale in producing X or Y the constraint will be non-linear over some range. In any case, if the community maximizes output Q, the familiar first-order condition is that the marginal rate of (technical) substitution between private and public inputs be equal to the ratio of their marginal costs:[17]

$$(3) \qquad \frac{\partial Q/\partial X}{\partial Q/\partial Y} = \frac{\partial B/\partial X}{\partial B/\partial Y}$$

In that $\partial Q/\partial X = (\partial Q/\partial XN^{1-\alpha})N^{1-\alpha}$, this formulation points up four important variables that contribute to the determination of the public-private choice of provision: relative costs of the inputs, the degree of "publicness" of the publicly provided good, the size or density of the community, and the productivity of each input for individual

[13]Tulkens and Jacquemin present a theoretical analysis in which individuals value two goods, one of which is subject to theft. The probability of theft is a function of public and private protection as well as the publicly provided pure public good of prevention. Although the level of security does not appear directly in their utility function, their analysis is similar to the present analysis in that they do note the importance of the marginal rate of substitution between public and private inputs.

[14]See Thomas Borcherding and Robert Deacon, p. 893, and Bergstrom and Goodman, p. 282. If, as is usually the case, benefits from publicly provided goods diminish with increasing distance as well as population, density would replace population as the measure of N.

[15]This aggregation of individual production relationships abstracts from the problem of differing tastes. However it would be consistent with a median voter model, for example, if household production functions were identical and linearly homogeneous. In that case, since the ratio x/y must be the same for all citizens, so must the level q_i; $q_i = f(XN^{-\alpha}, Y/N)$; $Nq_i = f(XN^{1-\alpha}, Y)$.

[16]The marginal product of police can correctly be measured in either of two ways: the marginal product of adding a police officer ($\partial Q/\partial X$) or the marginal product of an additional unit of police protection as viewed by citizens ($\partial Q/\partial XN^{1-\alpha}$). These are related as follows: $\partial Q/\partial X = (\partial Q/\partial XN^{1-\alpha})\ (\partial XN^{1-\alpha}/\partial X) = N^{1-\alpha}(\partial Q/\partial XN^{1-\alpha})$.

[17]In the case of police, $\partial B/\partial X$ is the cost, say, of an additional police officer, while $\partial B/\partial XN^{1-\alpha} = (\partial B/\partial X)/N^{1-\alpha}$ is the cost of an additional unit of police protection as seen by citizens. If $\alpha = 0$, if $N = 100$, and if one officer costs \$10,000, $\partial B/\partial X = \$10{,}000$ while $\partial B/\partial XN^{1-\alpha} = \$100$.

citizens. It is clear that the marginal products will not in general be independent of each other. Shoup suggests this in an example of protection against crime, noting that burglar alarms at some point may be more effective than further spending on police but that alarms alone would be of little use. His conclusion is consistent with the present analysis: "Layers of the marketing mode may thus be sandwiched in between layers of the group-consumption mode, on a least-cost basis" (p. 68).

The actual community allocation that results may differ from the optimal allocation described due to the nature of public goods and public decision making. One possible source of bias is the tendency for public goods to be undersupplied because of the widely speculated nonrevelation of preferences. While this tendency leads to an undersupply of public goods, by the same token it also tends to make private substitutes relatively more important. A second source of bias works in the opposite direction. If the institutions of revenue raising tend to produce "fiscal illusion" by making public provision appear to be less costly to voters than it actually is, a wedge is driven between the true budget line and the perceived budget line, causing a shift toward public provision.[18] Thus factors which have been discussed as affecting the level of public spending also naturally have implications concerning the relative importance of public and private provision, and must be considered along with the relative costs of the two modes of provision.[19]

This simple model of public and private provision demonstrates that the public-private split depends on the relative costs of inputs, the technology for producing the desired output, the public good character of the public input, and the institutions of public decision making and revenue raising. In Section III, several possible differences in technology between communities are considered, along with the effects of relative costs and public institutions. The relative importance of private provision will also be affected by the level of demand if parallel shifts in the budget lines alter the relative share of private provision, that is, if the technology implies nonhomothetic isoquants.

III. Substitution Between Public and Private Protection

In the production of security a large number of public and private inputs are involved. Public inputs include police, the courts, and prisons, while private inputs include locks, alarm systems, guards, protective agencies, and individual self-protective devices. Because comparable data on most of these inputs are not available, employment in two categories is analyzed: 1) public state and local police; and 2) private protective and detective agencies. The first category includes all police employees, not just police officers. The latter includes all employees of firms engaged in providing protective and detective services to households and firms (SIC 7393). The services of such firms range from providing security guards to installing and monitoring alarm systems. It is difficult to determine exactly what proportion of the business of such firms is carried out directly with households, but firms undoubtedly account for more revenue than households.[20] This does

[18]For a fuller discussion of fiscal illusion, see James Buchanan, ch. 10.

[19]A third source of bias exists if for some reason there is a tendency for the private substitute to be oversupplied. An example serves to demonstrate this point. One private means of providing protection against crime may be staying inside or taking a taxi rather than walking out at night. Because walking out at night probably produces external benefits in that such action makes streets safer for others, it will tend to be undersupplied. Put differently, the private alternatives of staying inside or taking taxis will tend to be oversupplied.

[20]Kakalik and Wildhorn, p. 63, present estimates indicating that over 85 percent of sales of private security equipment and services in 1968 were accounted for by firms and that only 1.6 percent of sales went directly to "consumers." By including equipment sales, however, this estimate probably overstates the share of firms compared to the employment data used in the analysis in this section. For an analysis of firm demand for protection, see Ann Bartel.

not alter the implications of the analysis as they apply to economy-wide substitution between private and public modes of producing protection.

Due to the practical as well as theoretical importance of substitution between privately and publicly provided inputs, a *CES* technology is assumed for the aggregation of labor inputs in protection. In the simplest case, nonlabor inputs are ignored and the production function is assumed identical for all communities. This yields a statement of how private labor inputs (Y) and the effective level of public labor inputs ($XN^{1-\alpha}$) are aggregated to produce the level of protection (Q) which is desired by households and firms.

$$(4) \quad Q = [A(XN^{1-\alpha})^{-\rho} + BY^{-\rho}]^{-1/\rho}$$

The technology as given implies that the elasticity of substitution between public and private inputs is $\sigma = 1/(1+\rho)$. Assuming that average factor wages equal the marginal costs of hiring those factors (in other words that the community budget constraint is linear), the optimal community allocation condition (3) may be rewritten as

$$(5) \qquad \frac{AN^{-\rho(1-\alpha)}X^{-\rho-1}}{BY^{-\rho-1}} = \frac{W_x}{W_y}$$

where W_x and W_y are the wages of the public input X and the private input Y, respectively. Taking *logs* and adding an error term u yields the equation

$$(6) \quad ln\ \frac{X}{Y} = a + \frac{1}{1+\rho}\ ln\ \frac{W_y}{W_x} - \frac{\rho(1-\alpha)}{1+\rho}\ ln\, N + u$$

where $a = [1/(1+\rho)]\,ln\,(A/B)$. Equation (6) may be appropriately estimated by ordinary least squares if the wage rate of each input is exogenously determined. The greater the elasticity of substitution $\sigma = 1/(1+\rho)$, the more nearly substitutable are public and private inputs. If $\sigma > 1$, an increase in the wage of one input will result in a decreased share of expenditures for that input. Community size matters only if $\alpha \neq 1$, that is, if the public input is not a strictly private good. The coefficient of $ln\ N$ may be positive or negative; it is positive if $\sigma > 1$ and $\alpha < 1$.

A. *Distortions in Community Choice*

Two sources of public-private bias are noted in Section II, however, which suggest that the efficient point described in equation (3) may not be reached. That is, the publicly provided input may be undersupplied because of nonrevelation of preferences or it may be oversupplied because voters are subject to fiscal illusion in their perception of the costs of government programs, causing them to vote for higher public spending. These effects would tend to distort the equilibrium condition for the allocation of resources between public and private inputs away from the efficient point described in (3). Since there is neither a theory to explain how the tendency towards nonrevelation of preferences for public goods might vary across states nor a reasonable proxy to measure such a tendency, its effect is not examined in this empirical section. The effect of the bias due to fiscal illusion may however be incorporated into the allocation decision reached by communities by rewriting the equilibrium condition as

$$(7) \qquad \delta\,\frac{\partial Q/\partial X}{\partial Q/\partial Y} = \frac{W_x}{W_y}$$

If fiscal illusion in fact creates no bias in the public-private choice, δ has a value of 1. If however fiscal illusion has the postulated effect of causing an oversupply of publicly provided goods, δ would have a value greater than 1, that is, $(\partial Q/\partial X)/(\partial Q/\partial Y) < W_x/W_y$.

Substituting the *CES* function into the modified equilibrium condition (7) yields

$$(8) \quad ln\ \frac{X}{Y} = a' + \frac{1}{1+\rho}\ ln\ \frac{W_y}{W_x} - \rho\,\frac{(1-\alpha)}{1+\rho}\ ln\, N + \frac{1}{1+\rho}\ ln\,\delta$$

Although there is no direct measure for the degree of fiscal illusion suffered by voters

in different states, one factor is in theory associated with the degree of fiscal illusion. It is the relative simplicity or complexity of the state's tax structure. Simpler tax structures are said to allow voters to perceive costs of public services more clearly, resulting in lower levels of spending.[21] If TS is a measure of tax structure simplicity, the relationship between fiscal illusion and tax simplicity is assumed to be

$$(9) \qquad \delta = TS^{-\theta} \qquad \theta > 0$$

That is, fiscal illusion is a decreasing function of tax simplicity. By substituting for δ, equation (8) may be rewritten

$$(10) \qquad ln\ \frac{X}{Y} = C + \frac{1}{1+\rho}\ ln\ \frac{W_y}{W_x} - \frac{\rho(1-\alpha)}{1+\rho}\ ln\ N - \frac{\theta}{1+\rho}\ ln\ TS + e$$

where C is a constant and e is an error term.

The measure of tax simplicity used in the estimation below is that developed by Wagner:

$$(11) \qquad TS = \sum_{i=1}^{6} (T_i)^2$$

where T_i is the proportion of a state's tax revenues which are collected by tax i.[22] This measure has a maximum value of 1, representing the simplest of all tax structures: total reliance on a single tax to raise all state revenue. The greater the reliance on a variety of taxes, according to the theory of fiscal illusion, the greater will be the complexity of the tax structure and the less will be the ability of voters to perceive all the costs of government programs. Thus the lower the value of TS, the greater will be the extent of any bias toward too much government spending.

[21] For a theoretical discussion of this point as well as an empirical test of the hypothesis, see Richard Wagner.

[22] The taxes are: 1) general sales, use, or gross receipts; 2) selective sales and gross receipts; 3) licenses; 4) income; 5) property; and 6) other taxes. See Tax Foundation, p. 178.

B. *Variations in the Production Function*

A second complication in estimating the elasticity of substitution between public and private inputs arises if the production function for protection varies across communities. For example, the relative efficiency of public police in providing protection may vary with population density for reasons quite apart from the possible public good nature of police protection. For example, public police might be relatively less effective in more urbanized areas if apprehension is more difficult in dense areas or if featherbedding and other inefficient practices are more prevalent in big city forces. In terms of the production function for protection, these possibilities suggest reasons why the marginal rate of substitution between public and private protective employment might be a function of population density, an effect comparable to Hicks-nonneutral technical change. Either of the parameters A or B in (4) could be a function of density, thus implying the addition of *log* of density (DEN) to equation (10). Another factor which might affect the relative efficiency of public and private protection is the industrial composition of the community. Public and private protection employment may vary in their effectiveness in protecting firms in different industries. Accordingly, three measures of industrial composition were also added to the estimated equation: per capita employment in manufacturing (MPC), wholesaling (WPC), and finance, insurance, and real estate (FPC).[23] Finally, it is possible that the production of security is not homothetic as assumed above and that changes in the demand for security may affect the public-private input split. To test this possibility, a number of social and demographic variables used by Bergstrom and Goodman in their analysis of expenditures for public police are included in the equations below.

[23] Employment by industry is taken from U.S. Bureau of the Census (1971a), Table 1B.

C. Estimation

In order to estimate the elasticity of substitution between public and private protective employment, equations based on (10) were fitted using cross-section data for states in 1970. States were used as observations, because data on private protection are not available by jurisdiction. The necessity of using aggregate data could of course bias or obscure relationships which would be observed if data by jurisdiction were available. In addition, the interpretation of the population variable (N) is made difficult. Accordingly, population was replaced by population per standard area, or density, in order to account for differences in the "publicness" of police services. In order to measure the average density as it applies to communities' opportunities to share public services, the measure used for each state is the average of county densities weighted by county populations. Because of the possibility noted above that density may measure the relative effectiveness of police forces in addition to reflecting the technological opportunities for sharing, however, it is impossible to give an unambiguous interpretation for the coefficient of density. Public police employment (X) includes state and local employees on a full-time basis, but employment in private protective and detective agencies (Y) is available only on the basis of total number of employees.[24] These measures do not account for any variations in average quality of labor across states.[25] The wage of public police employees (W_x) is estimated by the average wage—total police payrolls divided by full-time equivalent employees. The comparable average wage of the protective industry is not used for the private wage, however, because the occupational structure of the industry probably varies across states. Instead, the wage faced by private protective firms in state i (W_{yi}) is estimated by a weighted average of occupational wages in that state W_{ij}:

$$W_{yi} = \sum_j p_j W_{ij} \tag{12}$$

where the weight given to the wage of the jth occupation is the estimated national proportion of workers in the industry 7393 who are in that occupation.[26]

Equations estimated for the fifty states in 1970 are presented in Table 1. These regressions were weighted by the square root of state population because of apparent heteroskedasticity in the unweighted regressions. Included as explanatory variables are the variables in (10), the three measures of industrial composition introduced above, and seven variables used in Bergstrom and Goodman's analysis of police expenditures. This last group includes determinants of the demand for security in general. The hypothesis that they all have zero coefficients can be rejected at the 99 percent level. This finding suggests that the assumption of strict homotheticity is inappropriate in this case. Equation (1.1) in the table is estimated by ordinary least squares, and it yields a point estimate for the elasticity of substitution of 0.58. That the estimate is positive is consistent with the notion that communities respond in the expected direction to changes

[24] Public police employment and payrolls are taken from U.S. Bureau of the Census (1971b), Tables 8 and 9. Employment in protective and detective agencies is taken from U.S. Bureau of the Census (1973), State Reports, Table 183. It should be noted that security personnel hired directly by firms are not included in these private protective and detective agency data.

[25] Such variations might arise, for example, if unionization pushed wages up in one sector and this wage increase allowed higher quality workers to be hired in that sector.

[26] Proportions are based on the 1973 *Current Population Survey*, U.S. Bureau of the Census (1973), State Reports, Table 175: Guards and watchmen comprise two-thirds of all workers in the industry. Wage data are based on median earnings for men working 50–52 weeks for major occupational categories and the detailed occupational category of guards and watchmen. The weights applied to occupation wages are: professional, technical, and kindred workers, .031; managers and administrators, except farm, .042; sales workers, .010; clerical and kindred workers, .062; craftsmen and kindred workers, .083; operatives, except transport, .042; transport equipment operatives, .010; guards and watchmen, .667; other service workers, .052.

TABLE 1—ESTIMATED EQUATIONS, UNITED STATES, 1970
(Dependent Variable: $ln(X/Y)$)

	Equation (1.1) Ordinary Least Squares	Equation (1.2) Instrumental Variables[a]
Intercept	4.90 (6.12)	10.98 (8.07)
$ln(W_y/W_x)$ [σ]	.58 (.58)	2.47 (1.42)
ln DEN	−.0418 (.0649)	−.0167 (.0759)
ln TS	−.138 (.300)	−.461 (.404)
ln MPC	−.342 (.222)	−.252 (.260)
ln WPC	−.532 (.390)	−.471 (.446)
ln FPC	−.242 (.451)	−.395 (.525)
ln SH	−.186 (.803)	−.336 (.921)
ln 00	−1.14 (.68)	−2.14 (1.02)
ln EPR	.753 (.855)	1.20 (1.02)
DPOP	−.0214 (.0105)	−.0104 (.0140)
ln P65	.333 (.335)	.703 (.455)
ln Y	−.129 (.523)	−.377 (.619)
ln B	−.082 (.057)	−.101 (.066)
α	.694 (.224)	1.010 (.056)
R^2	.904	.876

Note: Variables are defined as: (X/Y) = ratio of public police employees to private protective employees; (W_y/W_x) = ratio of private protective wages to public police wages (see text); *DEN* = weighted population density (see text); *TS* = measure of tax simplicity (see text); *MPC, WPC, FPC* = employment in manufacturing, wholesaling, and finance, insurance and real estate per capita, respectively; *SH* = percent in same house five years before; *00* = percent housing units owner occupied; *EPR* = total employment-population ratio; *DPOP* = percentage change in population, 1960-70; *P65* = percent 65 and over; *Y* = median family income; *B* = percent black. Regressions were weighted by the square root of state population. Standard errors in parentheses.

Source: U.S. Bureau of the Census (1971a, b; 1973), and Tax Foundation (1973).

[a]Instruments for $ln(W_y/W_x)$ include the right-hand side exogenous variables plus the *log* values of the unemployment rate, state population, and W_y.

in relative wages for protective employment, but the large standard error makes it impossible to reject either the hypothesis that it is zero or that it is one.[27] Population density has a negative but insignificant coefficient. Ignoring any systematic relationship between density and the relative efficiency in the public production of police services (X), the density variable implies a parameter of publicness for police of .694.[28] This compares with estimates near 1.0 for police services calculated by Borcherding and Deacon, and Bergstrom and Goodman. If such a systematic relationship does exist, however, the estimate of this parameter will be biased. The measure of tax simplicity has the expected negative coefficient, but its standard error is relatively large. The coefficients of each of the industrial composition variables are negative but insignificant; these signs appear to indicate that private protective firms are more effective than public police at protecting firms in these industries. Of the variables indicating the demand for protective services, only the percentage change in population is significant by itself; the sign indicates that private protection is more effective or more readily responsive in areas of most rapid population growth.

One necessary assumption in order to use

[27]For comparison, the estimate of σ in the comparable equation omitting the Bergstrom and Goodman variables is 1.34, with a standard error of 0.56.

[28]Where c is the coefficient of *ln DEN*, the approximate mean and variance of α are:

$$E(\alpha) \approx 1 + \frac{c}{1-\sigma} + \frac{cov(c,\sigma)}{(1-\sigma)^2} + \frac{c\,var(\sigma)}{(1-\sigma)^3}$$

$$var(\alpha) \approx \frac{var(c)}{(1-\sigma)^2} + \frac{2c\,cov(c,\sigma)}{(1-\sigma)^3} + \frac{c^2\,var(\sigma)}{(1-\sigma)^4}$$

See D. V. Lindley, p. 135.

ordinary least squares in estimating equation (1.1) is that the wages for public and private protective employment are exogenously determined. Although this seems to be a reasonable assumption for the measure of private wages used—a weighted average of average occupational wages—it is less acceptable for the case of public police. Accordingly, an instrumental variables approach was used to estimate the same equation, letting the ratio of wages be treated as endogenous. The resulting estimates are shown in equation (1.2). Clearly, the point estimate of the elasticity of substitution is much larger (2.47), indicating that the technology of production on the local level may allow considerable latitude in responding to relative input prices. However, the standard error is relatively large, and it is possible to say only that this coefficient is significantly different from zero at the 90 percent level. The estimate of α is larger and quite close to 1.0. Other coefficients are quite similar to those in equation (1.1). The percent owner-occupied housing has a negative and significant effect on relative demand for public police, which is consistent with the results of Bergstrom and Goodman.

It is clear that these equations do not yield definitive estimates of the elasticity of substitution between public and private protective employment. Lacking better data, it is possible only to note the importance of this parameter. If there were any prior expectations as to the size of this elasticity, it seems at first glance that they would most likely point to a low elasticity, given that private and public police have traditionally had fairly well-defined legal and institutional responsibilities. Important functions such as arrest are reserved for public police. On the other hand, possibilities for substitution do exist for other functions such as surveillance, guarding, and some maintenance of order.[29] These opportunities for substitution illustrate why the community's demand for public and private protection may be quite sensitive to relative wage rates, and could therefore explain high estimates for the elasticity of substitution.

These possibilities for substitution may have great practical significance for the future of local government finances. If public sector unions are able to raise real wages faster than productivity increases or if the productivity of private inputs rises faster than private wages, communities will tend to substitute privately produced goods for public services. If the elasticity of substitution between public and private input exceeds one, the demand for private substitutes will increase proportionally more than public wage increases, thus decreasing the relative share of public spending for a given activity. This then could be a consequence of the increasing unionization within the public sector especially evident among police, teachers, and sanitation workers.

In the case of protection against crime, it seems clear that the costs of public inputs have increased relative to those of private inputs. For example, during the period 1967-73 the average salary for state and local police increased 56 percent. During the same period, the average salary for employees of private protective agencies increased 34 percent, the wholesale prices of door locks and padlock components had increases of from 15 to 35 percent, and the wholesale price of small arms increased 23 percent.[30] Depending on how substitutable these forms of private protection are with public police, price trends such as these tend to result in the substitution of private for public inputs. It is quite conceivable that this substitution could even result in an increase in aggregate crime rates if private means of protection work by simply diverting crime from protected to unprotected individuals instead of reducing aggregate crime rates. It should be emphasized, however, that the evidence presented in this sec-

[29]Bartel notes that guards who protect against shoplifting and employee theft are not substitutable while a protective service providing outside patrols "acts in a very similar capacity to that of public policemen ..." (p. 467). In her regressions explaining firm demand for protective workers, the number of police per capita exerts a generally negative but insignificant effect.

[30]Average salaries are payrolls divided by employment for the U.S. Public employment is full-time equivalent. *County Business Patterns* and *Public Employment*, various years. Wholesale prices are taken from U.S. Bureau of Labor Statistics (1974).

tion applies to only one form of private protection—employment in private protection agencies—and that the degree of substitutability may vary for other means.

IV. Conclusion

This paper presents a general analysis of private substitutes for public services and an application of this analysis to protection against crime. A central argument of the paper is that the extent of public provision of some services depends in part on the relative costs of publicly and privately provided inputs. These relative costs may be affected by institutions of public decision making, relative union strength, and relative efficiency of administration in the public and private sectors, as well as the technology of producing inputs in both sectors. One obvious implication is that the successful attempt by public employee unions to raise public sector wages without concomitant increases in productivity could result in relative or even absolute decreases in public employment, but failing that certainly a moderation in the rate of growth in local public spending. For the case of protection against crime, such opportunities for substitution may well provide one explanation for the recent growth of private means of protection in the United States. Whether this degree of substitutability extends to other forms of private protection besides protective employment or indeed to other classes of local services, however, remains a question for empirical investigation.

REFERENCES

A. P. Bartel, "An Analysis of Firm Demand for Protection Against Crime," *J. Legal Stud.*, June 1975, *4*, 443–78.

Y. Barzel, "Private Schools and Public School Finance," *J. Polit. Econ.*, Jan./Feb. 1973, *81*, 174–86.

W. J. Baumol, "Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis," *Amer. Econ. Rev.*, June 1967, *57*, 415–26.

G. S. Becker, "Crime and Punishment: An Economic Approach," *J. Polit. Econ.*, Mar./Apr. 1968, *76*, 526–36.

T. C. Bergstrom and R. P. Goodman, "Private Demands for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280–96.

T. E. Borcherding and R. T. Deacon, "The Demand for the Services of Non-Federal Governments," *Amer. Econ. Rev.*, Dec. 1972, *62*, 891–901.

D. F. Bradford et al., "The Rising Cost of Local Public Services: Some Evidence and Reflections," *Nat. Tax J.*, June 1969, *22*, 185–202.

James M. Buchanan, *Public Finance in Democratic Process: Fiscal Institutions and Individual Choice*, Chapel Hill 1967.

I. Ehrlich and G. S. Becker, "Market Insurance, Self-Insurance, and Self-Protection," *J. Polit. Econ.*, July/Aug. 1972, *80*, 623–48.

J. G. Head and C. S. Shoup, "Public Goods, Private Goods, and Ambiguous Goods," *Econ. J.*, Sept. 1969, *79*, 567–72.

James S. Kakalik and Sorrel Wildhorn, *The Private Police Industry: Its Nature and Extent*, Santa Monica 1971.

Michael Katz, *The Irony of Early School Reform*, Cambridge, Mass. 1968.

A. H. Lerman, "Selected Tables from A Survey of Charitable Contributions by Individuals," Office of Tax Analysis, Washington 1974.

D. V. Lindley, *Introduction to Probability and Statistics*, Part I, Cambridge 1965.

Blake McKelvey, *The Urbanization of America*, New Brunswick 1963.

M. V. Pauly, "Mixed Public and Private Financing of Education: Efficiency and Feasibility," *Amer. Econ. Rev.*, Mar. 1967, *57*, 120–30.

S. Peltzman, "The Effect of Government Subsidies-in-Kind on Private Expenditures: The Case of Higher Education," *J. Polit. Econ.*, Jan./Feb. 1973, *81*, 1–27.

Carl S. Shoup, *Public Finance*, Chicago 1969.

R. Stuart, "Billions for Protection," *New York Times*, Mar. 30, 1975.

J. Stiglitz, "The Demand for Education in Public and Private School Systems," *J. Publ. Econ.*, Nov. 1974, *3*, 349–85.

H. Tulkens and A. Jacquemin, "The Cost of Delinquency: A Problem of Optimal Allocation of Private and Public Expenditures," CORE disc. paper no. 7133, Catholic Univ. Louvain 1971.

R. E. Wagner, "Revenue Structure, Fiscal Illusion, and Budgetary Choice," *Publ. Choice*, Spring 1976, *25*, 45–61.

Tax Foundation, *Facts and Figures on Government Finance 1973*, New York 1973.

U.S. Bureau of Labor Statistics (*BLS*), *Survey of Consumer Expenditures 1960–61*, Washington 1966.

———, *Wholesale Prices and Price Indices, Supplement 1974*, Washington 1974.

U.S. Bureau of the Census, *U.S. Census of Population: 1970, Characteristics of the Population*, Washington 1973.

———, (1971a) *County Business Patterns, 1970*, Washington 1971.

———, (1971b) *Public Employment in 1970*, Washington 1971.

———, *Statistical Abstract of the United States*, 93d ed., Washington 1972.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, Feb. 1974.

U.S. News and World Report, "Private Police Forces in Growing Demand," Jan. 29, 1973, 54–56.

U.S. Office of Education, *Digest of Educational Statistics, 1973*, Washington 1974.

Reswitching, Wicksell Effects, and the Neoclassical Production Function

By DAVID LAIBMAN AND EDWARD J. NELL*

The Cambridge debate over capital theory has raised doubts about the validity of the neoclassical theory of distribution (see G. C. Harcourt). As this theory is widely assumed in empirical work, and often drawn upon in the analysis of policy, a demonstration that it is seriously flawed would require extensive rethinking of many areas of mainstream economics. Accordingly, two principal lines of defense have been advanced. The first, an oblique defense, accepts the critique, but asserts that only a simplifying "parable" has been damaged. The main corpus of neoclassicism, general equilibrium theory, remains unscathed. The second is a counterattack, and contends that a well-ordered neoclassical production function can be constructed after all (see Gallaway and Shukla, 1974, and refutations by Garegnani and by Sato, 1976).

In this paper we explore further dimensions of the problem by returning to the classical work of John Bates Clark to inquire into the issues of reswitching of techniques and negative price Wicksell effects (rises in the value of capital per man as the profit rate rises within a given technique) in a variety of settings. In the first section, following Clark, we set forth the problem of establishing the relation between the fund of capital as held in portfolios, and the heterogeneous capital goods used in production. The second section presents a two-sector indecomposable model with only circulating capital and profit formed on the wage as well as on capital-good inputs in primal (price) and dual (quantity) form. It is the model best adapted to the task of verifying the Clarkian story weaving behavior concerning capital funds and constraints concerning capital goods into an operational demand curve for "capital." In the third section, the reswitching possibilities are examined, and a special case identified for which a "nonreswitching theorem" does in fact hold. The fourth section examines the problem of negative price Wicksell effects and their impact on the attempt to construct a "surrogate" production function. The concluding section states results; the core of the matter, however, is that the capital controversies, against the rigorous Clarkian background, demonstrate the ill-founded character of both the aggregate production function and the general equilibrium defenses, and open the way for new approaches to distribution drawing on the post-Keynesian and Marxian heritage.

*Assistant professor of economics, Brooklyn College, and professor of economics, Graduate Faculty, New School for Social Research, respectively. This paper, in every way a joint product, originated in discussions in Nell's Theory Seminar at the New School and we would like to thank members of the seminar for constructive criticism of an earlier draft, which was written as a criticism of Lowell Gallaway and Vishwa Shukla (1974) and submitted to this *Review* at the same time (January 1975) as the comments by P. Garegnani and Kazuo Sato (1976). In revising it we have tried to move beyond our critical objective to interpret the production function controversy more generally. Comments by Luigi L. Pasinetti, Ian Steedman, and G. C. Harcourt are gratefully acknowledged, as are the suggestions of a referee of this journal.

I. Capital Funds and Capital Goods

Reswitching of techniques and perverse changes in the value of capital arise in the connection between two equally significant and practical concepts of capital—the portfolio concept, a homogeneous fund of value, shifted about in pursuit of the highest rate of return, and the productive concept, heterogeneous goods and equipment designed for specific uses. This is how the

problem of capital theory was first posed by Clark.[1]

The fund of capital moves in response to profit, and its effect is to produce a uniform rate of profit: "Given a certain permanent fund of capital... it is put into such forms that the rent secured by one concrete form or capital-good, is as large a fraction of its value as is that secured by another" (Clark, p. 125). But Clark did not think that capital goods themselves would be shifted around. Capital proper is mobile, capital goods are fixed. He wrote: "Capital is... the subject of competition; but capital-goods are not. The capital that is competed for does not consist in instruments—concrete, visible, moveable and ready for any one of a dozen different uses; there is no stock of capital-goods that has such adaptability..." (pp. 256–57).

In short, for Clark the problem of capital theory is that new capital is supplied and existing capital held in response to the rate of return on *funds*, while entrepreneurs, also necessarily motivated by the rate of return on the value of their investments, always demand capital for the purpose of embodying it in particular capital *goods*. What is required is a systematic connection between capital funds and capital goods, consistent with the relationship of each to the rate of profit. Clark's answer was that the more capital intensive a project, in terms of funds, the less an additional increment of funds could contribute to increasing the returns obtainable from it, by improving or altering the capital goods. Better technical ideas will be used first; it becomes progressively more difficult, hence more costly, to make technical improvements. Thus additional infusions of funds to improve capital goods will yield progressively smaller increments of profit. But the rate of return on the whole capital will be set, because of competition, by that on the final increment. Hence, with a given labor force, increasing capital funds can be embodied in alternative sets of quite different capital goods, which can be clearly ranked according to the associated rate of profit; the more expensive, more capital intensive the capital goods, the lower the rate of profit.

Clark thus does *not* put forth a simplistic "aggregate capital" theory of marginal productivity, or "parable." The fund of capital is a reality, significant in determining business behavior. On the other hand, the demand for capital can only be understood in terms of concrete heterogeneous capital goods. Economists who confine their analysis to heterogeneous capital goods are unable to deal with the supply side or with financial matters, while those who deal only in aggregate capital are incapable of analyzing the specifics of production and technical choice, and the corresponding rental values of particular capital goods. For Clark, neoclassical capital theory must encompass both funds and goods, and the connection between them must be consistent with the supply and demand framework. This is the point of constructing the aggregate production function, exactly as Paul Samuelson did, and it is this which the Cambridge critics contend they have shown to be impossible, in general.

II. The Two-Sector Model

In reconsidering the condition for successful coordination of capital funds and capital goods within a supply and demand framework, the appropriate model is a genuinely circulating capital model, in which the rapid turnover of capital goods

[1]Clark states: "It is inevitable that both capital and capital goods should be subjects of economic study. There are problems concerning each of them that have to be solved; and this fact appears in an unfortunate way, in all those treatises on political economy in which the single term, capital, is used to designate productive wealth. Invariably does the application of this term shift from capital, as we define it, to capital-goods, and vice versa. This two-fold meaning of one important word has made endless trouble and confusion....

The early economists all defined capital as consisting in instruments of production, such as tools, buildings, raw materials, etc.... Yet, having defined capital in this way, they were forced—as anyone must be—to revert to the common conception of it as a fund describable in terms of money, when they entered on the consideration of interest..." (p. 122).

See Hicks (1963), "Commentary," pp. 342–44 for a similar distinction.

releases funds that can migrate within a time frame consistent with a given set of conditions in financial markets. Thus, capital funds can "travel" from one set of capital goods to another in response to profit incentives, as in Clark's example of New England capital leaving whaling for textiles.[2] If capital goods turn over in the same time span as wage capital—the given production-payment period—profit must be formed on the capital advanced as wages as well as that advanced on materials.[3]

The price equations for an indecomposable two-sector model, incorporating the above assumption, are

$$(1) \quad 1 = (1 + r)(a_{01}w + m_1 + n_1 p_2)$$

$$(2) \quad p_2 = (1 + r)(a_{02}w + m_2 + n_2 p_2)$$

where r is the rate of profit, w the wage rate, p_2 the price of commodity two, m_1 and m_2 the input coefficients of commodity one into commodities one and two, respectively, and n_1 and n_2 the input coefficients of commodity two into commodities one and two, respectively. (Commodity one is the numeraire.) The price equations yield a wage-profit contour:[4]

$$(3) \quad w = \{1 - (m_1 + n_2)(1 + r) + (m_1 n_2 - m_2 n_1)(1 + r)^2\} + \{a_{01}(1 + r) + (a_{02}n_1 - a_{01}n_2)(1 + r)^2\}$$

and an expression for the price of commodity two:

$$(4) \quad p_2 = \frac{a_{02} + (a_{01}m_2 - a_{02}m_1)(1 + r)}{a_{01} + (a_{02}n_1 - a_{01}n_2)\ (1 + r)}$$

We now write down the quantity equations, dual to the price equations (1) and (2). These show the composition of output per head and the disposition of the labor force. Letting T_1 be output per head of commodity one, T_2 be output per head of commodity two, c can be consumption per head, where consumption is limited to commodity one (the numeraire),[5] and g is the uniform (steady-state) rate of growth, we have:

$$(5) \quad T_1 = (1 + g)(m_1 T_1 + m_2 T_2 + c)$$

$$(6) \quad T_2 = (1 + g)(n_1 T_1 + n_2 T_2)$$

$$(7) \quad 1 = a_{01} T_1 + a_{02} T_2$$

Eliminating T_1 and T_2, we find the consumption-growth curve:

$$(8) \quad c = \{1 - (m_1 + n_2)(1 + g) + (m_1 n_2 - m_2 n_1)(1 + g)^2\} + \{a_{01}(1 + g) + (a_{02}n_1 - a_{01}n_2)(1 + g)^2\}$$

which is parametrically identical to the wage-profit contour (3).

The most common approach to choice of technique assumes that the techniques compared have one process in common; goods are defined independently of the uses to which they are put, so that a sector's choice of technique is not affected by, nor does it affect, the choice of technique by any other sector. Further assuming that the numeraire

[2] Clark states: "As the vessels were worn out, the part of their earnings that might have been used to build more vessels was actually used to build mills. The nautical form of the capital perished, but the capital survived and, as it were, migrated from the one set of material bodies to the other" (p. 118).

[3] The assumption that profit is formed on materials alone dominates the circulating capital models in the literature; see Piero Sraffa, Hicks (1965), Samuelson, and Michael Bruno et al. While this procedure is most congenial to other aspects of the neoclassical vision (see Section IV), its justification is that circulating capital is conceived as a simple surrogate for fixed capital stocks. The immobility of fixed capital, however, makes the capital market adjustment story appear extremely dubious, as movements of capital can only be in response to uncertainly held expectations of future rates of return.

[4] Suppose commodity two were taken as numeraire; then the price equations would be

$$p_1 = (1 + r)(a_{01}w + m_1 p_1 + n_1)$$

$$1 = (1 + r)(a_{02} + m_2 p_1 + n_2)$$

with the wage-profit contour

$$w = \{1 - (m_1 + n_2)(1 + r) + (m_1 n_2 - m_2 n_1)(1 + r)^2\} + \{a_{02}(1 + r) + (a_{01}m_2 - a_{02}m_1)(1 + r)^2\}$$

[5] This single departure from full generality makes consumption per head, in physical units, dimensionally homogeneous with the wage rate. If both commodities were consumed, consumption per head would have to be evaluated using the price ratio, and the consumption-growth tradeoff would not be independent of, let alone dual to, the wage-profit tradeoff.

sector (sector one) is the one with the technique held in common, use the following compact notation for the case of two techniques:

$$\begin{aligned} (9) \qquad s &= s_1 = a_{01} \\ t &= (a_{02}n_1 - a_{01}n_2) \\ t_1 &= (a_{02}n_1 - a_{01}n_2)_1 \\ j &= (m_1 + n_2) \\ j_1 &= (m_1 + n_2)_1 \\ k &= (n_1 m_2 - m_1 n_2) \\ k_1 &= (n_1 m_2 - m_1 n_2)_1 \end{aligned}$$

The subscript 1 applied to the t, j, and k refers to the second technique. Equating the expression for the wage in the two techniques ($w = w_1$), we get the switching equation:[6]

$$(10) \quad (1+r)^2 + \frac{s(k_1 - k) + tj_1 - t_1 j}{tk_1 - t_1 k} \cdot (1+r) + \frac{s(j_1 - j) + (t_1 - t)}{tk_1 - t_1 k} = 0$$

For reswitching of techniques to occur, (10) must have two positive roots for which $0 \leq r \leq r_0$, where r_0 is the maximum rate of profit at which either w or w_1 becomes zero. Sato neatly derives the range of values for the technical coefficients within which this condition is fulfilled.[7]

III. Varieties of Reswitching

Reswitching of techniques—in which a single technique is the most profitable at both high and low rates of profit—is the impasse that brings the Clark expedition to a halt. We examine it under a variety of assumptions.

A. *The Gallaway and Shukla Assumption*

We add to the record our counterexample to the "general nonreswitching theorem" proposed for the case in which all $k_i, t_i > 0$:

Parameter	Technique A	Technique B
a_{01}	1.0	1.0
m_1	0.1	0 1
n_1	1 0	1.0
a_{02}	0.55872	0.56712
m_2	0.135872	0.261712
n_2	0.35872	0.11712

Using (9), we have $s = s_1 = 1$, $t = .20$, $t_1 = .45$, $j = .45872$, $j_1 = .21712$, $k = .10$, $k_1 =$

[6]When profit is not calculated on the wage, the wage-profit curve will be

$$w = \{1 - (m_1 + n_2)(1+r) + (m_1 n_2 - m_2 n_1)(1+r)^2\} \div \{a_{01} + (a_{02}n_1 - a_{01}n_2)(1+r)\}$$

Comparing this with (3), we see that the only difference consists in a factor of $(1+r)$ in the denominator. When, solving for switch points, we set $w = w_1$, the factor $(1+r)$ will cancel out. Hence, given the technologies, switch points are the same whether or not profit is figured on the wage.

[7]Gallaway and Shukla (1974) propose a restriction on the coefficients sufficient to render commodity prices "positive and finite for all positive values of r." This seemingly reasonable restriction is in fact excessive; all that is necessary for an economically reasonable model is that prices (and the wage) be positive for $0 \leq r \leq r_0$ (where r_0 is the maximum rate of profit), which is guaranteed by the well-known Perron-Frobenius Theorem—the maximal eigenvalue of a nonsingular matrix, and it alone, has an all-positive eigenvector associated with it. A rigorous but nontechnical presentation will be found in ch. 5 of Sraffa. A simple demonstration for nonmathematicians: rewrite the numerator of (4) as

$$(4') \qquad a_{02}[1 - m_1(1+r)] + a_{01}m_2(1+r)$$

Now consider a *comparison system* with $n_1 = n_2 = 0$. This system will have a maximum rate of profit $\hat{r}_0$ greater than r_0 of the system given by (1) and (2). For the comparison system, setting $w = 0$ in (3), we find $1 + \hat{r}_0 = 1/m_1$. For this value of r, (4') reduces to $a_{01}m_2/m_1 > 0$, demonstrating that the numerator must be positive for $\hat{r}_0$, hence for r_0 and for all r such that $0 \leq r \leq r_0$. A symmetrical demonstration, this time with $m_1 = m_2 = 0$ in the comparison system, applies to the denominator of (4), which must therefore be positive for all relevant r in any productive technology. The "counterexamples" tabulated by Gallaway and Shukla are therefore all perfectly valid examples of reswitching, as prices are positive for profit rates lower than the maximum rate; see Garegnani (1976). The Gallaway and Shukla restriction applies to (4); it is $a_{01}m_2 - a_{02}m_1 > 0$, $a_{02}n_1 - a_{01}n_2 > 0$. This implies a very special and arbitrary ordering of the coefficients. Let $A = n_1/n_2$, $B = a_{01}/a_{02}$, $C = m_1/m_2$. The Gallaway and Shukla restriction implies $A > B > C$. Call this ordering ABC. The possible orderings then are: ABC, CAB, BCA, ACB, BAC, CBA. In short, Gallaway and Shukla focussed on one out of six possible configurations of the coefficients. For our purposes we note here that their ordering implies all $k_i, t_i > 0$.

.25. Substituting these values into (10) and simplifying yields:

$$(1 + r)^2 - 2.6(1 + r) + 1.68 = 0$$

which factors neatly into

$$[(1 + r) - 1.2][(1 + r) - 1.4] = 0$$

Thus there are two positive roots, $r_1 = .2$ and $r_2 = .4$. The maximum rate of profit for the first technique is $r_{0_1} = .6129$, for the second $r_{0_2} = .61176$; hence, both roots fall within the required range.[8] The Gallaway and Shukla examples join the ranks of the standard reswitching examples which violate the unjustified assumption $k_i, t_i > 0$ (see fn. 7).

B. *Principal Diagonal* = 0

It is possible to find an economically meaningful condition strong enough to yield the conclusion that the Clark story requires—but it is a very special case, indeed. The condition that

$$(11) \qquad j = j_1 = 0$$

gives the required result. For then the switching equation becomes

$$(10') \quad (1 + r)^2 + \frac{s(k_1 - k)}{k_1 t - k t_1}(1 + r) + \frac{t_1 - t}{k_1 t - k t_1} = 0$$

By definition, now, $k_i, t_i > 0$. For there to be two positive real roots the term associated with $(1 + r)$ must be negative and the constant term positive. First, suppose $k_1 t > k t_1$. Then $k_1 < k$ implies $t > t_1$, so the constant term will be negative. But if $k_1 > k$, the $(1 + r)$ term will be positive. Next suppose $k_1 t < k t_1$. If $k_1 < k$, the $(1 + r)$ term will be positive. But if $k_1 > k$, then $t_1 > t$, which will make the constant term negative. Hence, there cannot be two positive roots, and reswitching is ruled out.

This tells us at best that neoclassical theory can only deal with highly specialized and unusual cases of changes in technique. An adequate theory will have to range further afield, into regions where the marginal productivity principle can give no guidance. Reswitching is not a peculiar or "perverse" phenomenon; on the contrary, it is generally possible, and it is the neoclassical case that is special.

C. *Intersectoral Switching of Techniques*

Comparison of techniques with one process in common considerably limits the generality of the results, for in a multisector world, it means comparing economies which differ in one process only.

Consider the problem of choice of techniques as Joan Robinson originally posed it: there are a number of identical islands in a sunny archipelago, the engineers of which all read the same book of blueprints. For historical or extraeconomic reasons different wage rates prevail on different islands. Arrange the islands in order, according to their wage rates. The question of "reswitching" now carries no implication that there is any change from one technique to another and then back; rather we wish to know if an island with a low wage and one with a high wage might find the same technique most profitable, while an island with a middling wage chose a different one.

The assumption that the techniques differ in only one process carries the implication that producing the second good in a different way would nevertheless yield a product capable of entering just as it did before into

[8] The Gallaway and Shukla proof of their nonreswitching theorem relies on comparison of the ratio of the numerators of the wage-profit curves of the two techniques (N) with the ratio of the denominators (D). Their account of the denominator ratio curve is correct; it has a negative first and a positive second derivative, falling but flattening out and remaining positive throughout the range $0 \leq r \leq r_0$. But in their account of the N curve (p. 355, Figure 5, and p. 356, Figure 7), they fail to realize that it may cut the D curve from above, and then flatten out more rapidly (or begin to rise) and cut it again, within the relevant range. This is what happens in our reswitching example in the text, as shown by these computations, for selected values of r:

$r = 0$	$N = .8281$	$D = .8275$	$N > D$
$r = 0.2$	$N = .8051$	$D = .8051$	$N = D$
$r = 0.3$	$N = .79481$	$D = .79485$	$N \leq D$
$r = 0.4$	$N = .785276$	$D = .785276$	$N = D$
$r = 0.5$	$N = .7773207$	$D = .776119$	$N > D$

the first process. While this may often be the case, it is also commonly true that the specifications laid down by industrial consumers of a product limit or determine product design, materials, machine tooling, and the method of production in general. The modern neoclassical assumption that many different methods exist for producing exactly the same product simply ignores the inherent connection between the potential uses of a thing and the way it is made.[9] If a change in the method of production changes the qualities of the product, this may in turn require a change or adaptation in the method of production of the second good.[10]

The mathematical argument[11] that the total number of processes cannot exceed the total number of unknowns—the $n - 1$ prices, plus the wage and the rate of profits—is perfectly correct, but begs the question. For this argument assumes that the systems are in contact, and that the equations are to be solved simultaneously. Then only at switch points will prices be identical across techniques; hence, the price system will tell us which method is cheaper.[12] But in comparing island economies or more generally in choosing methods of production, we are interested not in prices but in profit rates and capital values, so the fact that the techniques are associated with different price structures will not affect the decision.

Reswitching examples in this case are readily available; see the one below with $k_i, t_i > 0$, as in Section IIIA.

Parameter	Technique A	Technique B
a_{01}	10.000	10.730
m_1	0.200	0.100
n_1	1.000	1.000
a_{02}	3.000	3.219
m_2	0.220	0.275
n_2	0.100	0.100

Here, technique A is used at low profit rates; there are switch points at $r = .09$ and $r = .32$, approximately, and technique A returns for r between .32 and $r_0(B) = .6085$.

D. *Profit Maximization*

If the world were coming to an end tomorrow, firms would maximize profits instead of the profit *rate*.[13] The rate of return becomes irrelevant, since growth is moot, and in the immediate future everything will be bygones. In the Clarkian world, and indeed the real world, however, there is a capital market; rental values of capital goods are ultimately governed by the rate of interest on capital funds. Efficiency shadow pricing of given endowments of resources, as in much contemporary general equilibrium theory, could not be acceptable to Clark. So long as the ratio of rental earnings to endowment (replacement) value remains nonuniform, capital funds will move, until the ratio of rents to value of holdings is equalized. Marginal product shadow prices are not, in general, equilibrium prices.

Suppose an entrepreneur invested in a method of production that maximized profits per head but yielded a lower rate of profit. He must borrow capital and pay the going rate of interest, while facing the same

[9]For a critique of this assumption, see Martin Hollis and Edward Nell, ch. 9, especially pp. 234-40, 245-48. As regards the assumption, see Clark. "This complementarity of producers goods must always be considered: since a poor machine introduced into an equipment of good ones has the effect of taking something from the productive power of the other parts of the equipment" (p. 248). For Clark, altering methods and improving products go hand in hand, and a method cannot be altered without its suppliers of inputs appropriately improving their products (pp. 246-52).

[10]Perhaps the best-known case of switching where both sectors always change methods together is Samuelson's construction of the surrogate production function. Since the two sectors always have the same capital-labor ratio, they will both cease to be most profitable at the same time. The techniques always have different price structures, even at switch point, with prices equal to labor values in each case.

[11]See Sraffa, pp. 81-82; Bharadwaj, p. 415.

[12]Moreover, this is necessary for the rate of profit to equal the real marginal product of capital at a switch point. Consider two techniques in aggregative terms: $y_1 = rk_1 + w$, $y_2 = rk_2 + w$. Clearly, $r = (y_2 - y_1)/(k_2 - k_1)$. This can be interpreted as a "marginal product" *only* if the two systems have identical price structures at the switch point. Moreover, it is an accounting marginal product, and cannot be used to determine the rate of profit.

[13]In their reply to Garegnani and Sato, Gallaway and Shukla revert to a profit-maximization assumption as a second line of defense.

prices and wage rates as all other producers. Profit rate maximizers will be able to pay a higher rate of interest, thus attracting the entire supply of capital to themselves; profit maximizers would prove unable to meet competitive costs.

Still, since profit maximization is closely related to the general equilibrium defense against Cambridge criticism, we consider the possibility of reswitching in the circulating capital model, in which the technique is chosen which maximizes profits per head (π). For the aggregate economy, we have the quasi identity

$$w(1 + r) + rk = c + g(w + k) \tag{12}$$

This gives net national product seen as wages plus profits on the left, consumption plus net investment on the right (remembering that in a model where wages are advanced, so must consumption be accumulated and advanced). Solving for the value of the total capital flow,[14]

$$w + k = (c - w)/(r - g) \tag{13}$$

The value of capital per head is then given by the slope of the chord connecting the wage-profit point and the consumption-growth point on the single curve representing both the wage-profit and the consumption-growth contours (equations (3) and (8)). The w or c intercept of this chord, then, gives the value of net output per head, and the r or g intercept, the ratio of the value of net output to the value of capital.

If g goes to zero, the w-intercept of the chord comes to coincide with the w-intercept of the contour itself. In this zero-growth case, the value of net output per head is invariant to the rate of profit, and the technique with the highest w-intercept will always produce the maximum profits per head, whatever the level of the wage.[15]

While the growth rate is irrelevant to the reswitching issue as conventionally defined, it is important in this case, because the value of profits per head unlike the profit rate is not invariant to the composition of output. To give our argument some rope, we assume a high growth rate, implying substantial savings out of wages (we are concerned with the possibility, not the empirical likelihood, of reswitching; examples using weaker assumptions can undoubtedly be found).

In Figure 1, two techniques are drawn with no conventional reswitching, and no negative price Wicksell effects (see the next section). The growth rate is g_0, the maximum rate for technique A. Consider three levels of the wage: at w_1, $\pi_A = h - w_1 > \pi_B = 0$; at w_2, $\pi_B = j - w_2 > \pi_A = k - w_2$; finally, at w_3, $\pi_A = m - w_3 > \pi_B = n - w_3$. Clearly, on the profit-maximization criterion, technique A is most profitable at the high- and low-wage levels, and technique B is most profitable at the intermediate level. The inescapable curse of reswitching reappears.

IV. Negative Price Wicksell Effects and the Surrogate Production Function

The neoclassical story, whether in rigorous-Clarkian or jelly-parable form, requires an inverse monotonic relationship between the rate of profit and fund- or real-capital, respectively. In the absence of reswitching and perverse shifts in the value of real capital at switchpoints (capital reversing), validation of the inverse relation would depend on the premise of infinite substitutability, unless it were shown that the value of capital varies inversely with the profit rate within a single technique. In view of the attention devoted to this condition and its relation to efforts to construct a surrogate production function, we will examine, within the circulating-capital two-sector model, the conditions necessary for absence of negative price Wicksell effects (increases in the value of capital per unit of labor as

[14]See Spaventa and Nell for the derivation of this relationship for models in which profit is formed on physical capital only.

[15]This is the Gallaway-Shukla case, which thus rests on the arbitrary restriction $g = 0$. Note that to rule out reswitching, these authors are ready to rule out *any* switching of techniques—and this in the name of a defense of neoclassicism!

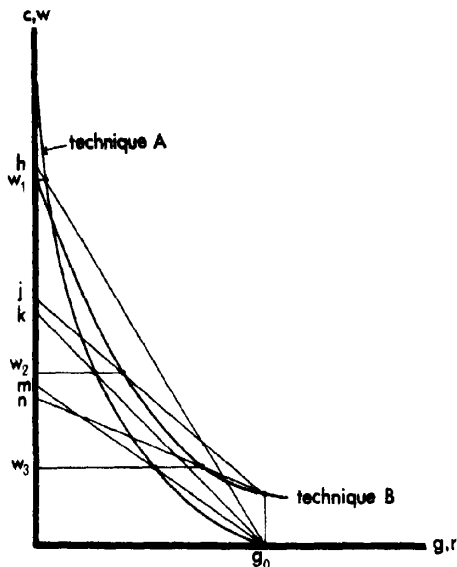



FIGURE 1

the profit rate rises).[16] In Figure 1, as noted above, the value of capital per man is the slope of the chord connecting the $c - g$ point and the $w - r$ point; if this slope is to rise with a rise in the profit rate, the contour must be concave from below, and the condition for negative price Wicksell effects is thereby identified: the wage-profit (or consumption-growth) contour must have a negative second derivative.

Gallaway and Shukla attribute the absence of negative price Wicksell effects to their assumption $k_i, t_i > 0$, which we have already shown to be without economic justification. We will now show that it is also necessary that profit be calculated on the wage. The two conditions are jointly sufficient, as Gallaway and Shukla demonstrate in their original article, but neither alone will suffice.[17]

Consider the following examples: First, suppose that the condition $k_i, t_i > 0$ holds, in the strongest possible form—$m_1 = n_2 = 0$—but that profit is not calculated on the wage. Let $a_{01} = 1, a_{02} = .2, n_1 = .1, m_2 = .8$. The wage-profit contour is

$$w = \frac{1 - .08(1 + r)^2}{1 + .02(1 + r)^2}$$

and $r_0 = 2.5355$, $w_0 = .902$. Let $R = (1 + r)$. Then

$$dw/dR = (-.2R)/(1 + .04R^2 + .0004R^4) < 0$$

$$d^2w/dR^2 = (-.2 + .008R^2 + .00024R^4)/(1 + .04R^2 + .0004R^4)^2$$

which is < 0 for $R \leq 3.5355$.

Now consider a case where profit is calculated on the wage, but where $t < 0$, $k = 0$. Let $a_{01} = 1$, $a_{02} = .2$, $m_1 = .1$, $m_2 = .7$, $n_1 = .1$, $n_2 = .7$. Then $k = 0$, $j = .8$, $t = -.68$, and $s = 1$. The maximum rate of profit is .25. Differentiating (3), we obtain

$$dw/dR = (-.544R^2 + 1.36R - 1) \div (R^2 - 1.36R^3 + .4624R^4)$$

which is negative between $R = 1$ and $R = 1.25$;

$$\frac{d^2w}{dR^2} = \{.5032R^4 - 2.627R^3 + 5.5488R^2 - 5.44R + 2\} \div \{.2138R^7 - 1.258R^6 + 2.774R^5 - 2.72R^4 + R^3\}$$

which is also negative in the range $1 \leq R \leq 1.25$.

Only when both the capital-intensity condition $k_i, t_i > 0$ holds, *and* profit is figured on the wage, will negative price Wicksell effects be eliminated.

[16]Gallaway and Shukla, in attempting to rule out this phenomenon by showing that $d^2w/dr^2 > 0$ for a single wage-profit curve (3), mistakenly argue that this is a necessary condition for a "factor price contour [which is] monotonically decreasing and decreasing at a decreasing rate throughout the positive quadrant" (p. 357). They seem to have forgotten that Samuelson's surrogate production function is made up of *linear* wage-profit tradeoffs.

[17]It is easy to see why figuring profit on the wage helps to rule out negative price Wicksell effects. There is a massive "composition effect" working against them. Since the rate of profit is formed on the wage, the wage is part of capital, and a rise in profits involves a fall in the capital flow equal to the fall in the wage. The rise in the price of physical capital inputs must be huge to offset this fall and ensure an overall rise in the value of capital.

This double condition suggests that the attempt to rule out valuation perversity ends up on the horns of a dilemma. First, we observe that the model in which profit is not formed on the wage is most appropriate as an analog to the real world of real capital stocks, fixed in form for long periods of time, compared to which the stocks of financial capital tied up in payrolls are negligible. With the value of capital dominated by stocks of physical capital, the possibility of "perverse" swings in value relative to profit rates is all the greater.

However, as explained above, we seek to give full play to the Clark story in which financial capital "migrates" from one physical body to another; we therefore return to the model with profit calculated on the wage, in which negative price Wicksell effects are less likely. But we now observe that the "well-ordered" neoclassical production function requires more than the nonreswitching of techniques. Marginal products must equal rates of factor payment. More generally, Samuelson's surrogate production function is constructed on the basis of a central and defining property of linear homogeneous production functions—that the elasticity of the wage-profit frontier measures relative shares.[18] But when the rate of profit is calculated on the wage these properties cannot hold consistently with marginal productivity theory. Starting with the net income equation, in obvious notation:

$$(14) \quad Y = wN + rwN + rK$$

Rewrite in per capita terms:

$$(15) \quad y = w + r(w + k)$$

Differentiating, and manipulating,

$$(16) \quad dy = dw + rdw + wdr + kdr + rdk$$

$$(17) \quad \frac{dy}{dk + dw} = r + \frac{dw}{dk + dw} + \frac{dr}{dk + dw}(k + w)$$

This implies that

$$(18) \quad \frac{dy}{dk + dw} = r \text{ if and only if } \frac{dw}{dr} = -(k + w)$$

The marginal net product of total capital equals the rate of profit if and only if the slope of the wage-profit curve equals the value of total capital (ignoring sign). There is, of course, no reason to suppose this condition will be met; but if it is, then the elasticity of the wage-profit curve measures relative shares:

$$(19) \quad -\frac{rdw}{wdr} = \frac{r(k + w)}{w}$$

But the reciprocal condition that the marginal product of labor equal the wage must also hold.[19] Rewriting the net income equation and using $\phi = Y/K$, $n = N/K$ in per unit of capital terms:

$$(20) \quad \phi = (1 + r)wn + r$$

we find, using the same method as before,

$$(21) \quad d\phi = (1 + r)ndw + (wn + 1)dr + (1 + r)wdn$$

$$(22) \quad \frac{d\phi}{dn} = (1 + r)w, \text{ if and only if } \frac{dw}{dr} = -\frac{wn + 1}{(1 + r)n}$$

[18]Write a linear homogenous "well-behaved" production function in per capita form. $y = y(k)$. Then $r = y'(k)$ and from Euler's Theorem, $w = y(k) - ky'(k)$; $dr/dk = y''(k)$; $dw/dk = y'(k) - y'(k) \cdot (dk/dk) - ky''(k)$. Hence $dw/dr = (dw/dk)/(dr/dk) = -k$, $rdw/wdr = -rk/w = -rK/wL$. The wage-profit frontier is derived from the cost-minimizing problem which is dual to the output maximizing usually associated with the production function; see Samuelson, p. 202, fn. 2.

[19]This is not explicitly recognized by Amit Bhaduri. Nevertheless it can be shown that the condition does indeed hold in his model which calculates profit on physical capital only.

which, rearranged, is equal to

$$\frac{dw}{dr} = -\frac{k + w}{1 + r} \tag{23}$$

Thus the marginal net product of labor will equal not the wage, but the wage *marked up by the gross profit rate*, if and only if the slope of the wage-profit curve equals the value of capital *discounted by the gross profit rate*. In the model with profit calculated on the wage, then, the two marginal productivity conditions cannot be simultaneously met. The model does not generate a surrogate for use in the neoclassical parable.

And thus the dilemma: With profit formed on advanced wages, negative price Wicksell effects *can* be ruled out, but the marginal productivity conditions fail; in the opposite case, the latter conditions and their surrogate properties are preserved—here the reswitching and capital-reversing problems are not even considered—but only by incurring the necessary possibility of negative price Wicksell effects.

A final remark: Calculating profit on the wage appears to abandon a crucial neoclassical position, one stated very forcefully by Clark, namely that labor and capital are separate, independent, and symmetrical factors, cooperating in production and receiving rewards commensurate with their contributions. To figure profit on the wage bill amounts to a return to one of the oldest conceptions of capital as, in part, a wage fund. This conception is alien to the structure of capital as it has evolved in modern industrial conditions, but it does undermine the "factors of production" approach; we therefore wonder whether it is consistent with the defense of neoclassical theory.

V. Conclusion

We have shown that the continuing attempts to rule out perverse phenomena in the neoclassical theory of production and distribution have validated the contention of the Cambridge critics: that the crucial connections between Clark's capital funds and capital goods cannot be made in the way required, except in special and implausible conditions.

But the central point is not whether reswitching, capital reversing, and counterindicated valuation shifts are empirically likely or unlikely.[20] The critics' deeper contention is rather that the supply and demand framework within which the problem of distribution is posed must be reassessed. One approach, mentioned earlier, is to abandon the fund concept of capital; this is the path taken by the general equilibrium theorists. This would have been just as unacceptable to Clark as the opposite reduction of capital to "aggregate" jelly, now largely seen to be inadequate. Equilibrium prices which assign rental values to capital goods but fail to provide a uniform rate of profit[21] on capital funds will not do, for they simply make impossible any theory of the central institution of the capital market.

It is not our purpose to argue the case against general equilibrium theory here. We do wish to assert what will seem to many a great irony: that the Cambridge critics of neoclassical theory are closer to Clark than either the general equilibrium enthusiasts or the would-be defenders of the parable, such as Gallaway and Shukla. The controversy has shown that, despite the vigor of Clark's original vision, it is not possible to consistently relate capital goods (in the theory of production) and capital funds (in the distribution of income to property ownership) within a comprehensive supply and demand framework that rules out social relations other than market relations grounded only in exogenously given technology and preferences. If anything is to be learned from this, it is to retain Clark's appreciation of the complexities of the capital concept in developing the institutionally richer post-Keynesian and Marxist theories of distribution.

[20]But see Peter Albin, who examines the case of a major lumber company switching from horses and men to mechanized methods of hauling logs with a rise in wages, and then back to horses and men with a further rise.

[21]Except with linear Engel curves, constant returns, and a very special set of initial conditions.

REFERENCES

P. Albin, "Reswitching: An Empirical Observation," *Kyklos*, 1975, No. 1, *28*, 149–54.

A. Bhaduri, "On the Significance of Recent Controversies on Capital Theory: A Marxian View," *Econ. J.*, Sept. 1969, *79*, 532–39.

K. Bharadwaj, "On the Maximum Number of Switches Between Two Production Systems," *Schweitz. Z. Volkswirt. Statist.*, Dec. 1970, *106*, 409–29.

M. Bruno et al, "Nature and Implications of the Reswitching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 526–53.

D. G. Champernowne, "The Production Function and the Theory of Capital: A Comment," *Rev. Econ. Stud.*, 1954, No. 2, *21*, 112–35.

John B. Clark, *The Distribution of Wealth: A Theory of Wages, Interest and Profits*, New York 1956.

L. Gallaway and V. Shukla, "The Neoclassical Production Function," *Amer. Econ. Rev.*, June 1974, *64*, 348–58.

——— **and** ———, "Reply," *Amer. Econ. Rev.*, June 1976, *66*, 433–36.

P. Garegnani, "Switching of Techniques," *Quart. J. Econ.*, Nov. 1966, *80*, 554–67.

———, "Heterogeneous Capital, the Production Function and the Theory of Distribution," *Rev. Econ. Stud.*, July 1970, *37*, 407–36.

———, "The Neoclassical Production Function: Comment," *Amer. Econ. Rev.*, June 1976, *66*, 424–27.

G. C. Harcourt, *Some Cambridge Controversies in the Theory of Capital*, London; New York 1972.

John Hicks, *Capital and Growth*, New York 1965.

———, *Theory of Wages*, 2d ed., London 1963.

Martin Hollis and Edward J. Nell, *Rational Economic Man: A Philosophical Critique of Neo-Classical Economics*, London; New York 1975.

E. J. Nell, "A Note on Cambridge Controversies in Capital Theory," *J. Econ. Lit.*, Mar. 1970, *8*, 41–44.

J. Robinson, "The Production Function and the Theory of Capital," *Rev. Econ. Stud.*, 1954, No. 2, *21*, 81–106.

———, "Capital Theory Up to Date," *Can. J. Econ.*, May 1970, *3*, 309–17.

P. A. Samuelson, "Parable and Realism in Capital Theory: The Surrogate Production Function," *Rev. Econ. Stud.*, June 1962, *29*, 193–206.

K. Sato, "The Neoclassical Production Function: Comment," *Amer. Econ. Rev.*, June 1976, *66*, 428–33.

Piero Sraffa, *Production of Commodities by Means of Commodities*, London; New York 1960.

L. Spaventa, "The Rate of Profit, Rate of Growth and Capital Intensity in a Simple Production Model," *Oxford Econ. Pap.*, July 1970, *22*, 129–47.

Distributional Neutrality and Optimal Commodity Taxation

By DAVID E. WILDASIN*

"We are still a long way from having an intuition for resource allocation questions in economies with distorting taxes which parallels the level of intuition in first-best economies."
Peter Diamond [p. 342]

The theory of optimal commodity taxation has been elaborated and refined in recent years by a long list of notable contributors. There are, however, certain areas that need extension and clarification. Much of the discussion of optimal tax rules has focused on the case of an economy with a single consumer or many identical consumers. The restriction to a single consumer is particularly serious, because certain results appearing with great regularity in the literature are not generally valid in the many-consumer case. The more careful studies emphasize this point.[1] Manifestly, consumers are not identical, and the question arises whether we have learned, or ever could learn, anything really interesting by studying optimal taxation under this assumption. It is clearly important to understand why the many-consumer case is different.

In this paper I shall examine in detail two tax rules which have appeared in a number of writings on optimal taxation. One well-known rule states that the ratio of the additional revenue obtained by a unit increase in the tax on a commodity to the quantity of the commodity should be the same for all commodities. Another rule, due to Ramsey, states that taxes should induce (approximately) equal percentage reductions in compensated demand for all commodities. Both of these rules are valid in the identical-consumer case; neither is ***generally*** valid in the many-consumer case. A major purpose of the discussion below is to study the implications of two restrictions that can be imposed on the social welfare function, each of which might be considered a formalization of the idea of "distributional neutrality." One of these (simple neutrality) states that the marginal social utility of *consumption* of the numeraire good (i.e., the

*Assistant professor of economics, University of Illinois at Chicago Circle. I am grateful to the managing editor and a referee for critical comments on an earlier and much longer version of this paper; neither is responsible for remaining errors.

[1]Among those who have explicitly assumed identical consumers are Frank Ramsey in his pioneering work, Avinash Dixit, Joseph Stiglitz and Partha Dasgupta, Anthony Atkinson and Stiglitz (1972), Agnar Sandmo (1974), and Atkinson and Nicholas Stern. William Baumol and David Bradford are not explicit on the point, but their results in fact hold (generally) only in the case of identical consumers. Dasgupta and Stiglitz, James Mirrlees (1972), and Frank Hahn discuss the case of nonidentical consumers, but only in regard to the problem of profit taxation and production efficiency—matters which I will not discuss here. The many-person commodity tax problem was treated by Marcel Boiteux in his classic 1956 paper. Boiteux, it should be noted, did assume that income can be costlessly redistributed in lump sum fashion. This work was introduced to English-language economists by Jacques Drèze, who explicitly draws attention to the role of redistributive transfers in the Boiteux analysis. This point was not lost on Herbert Mohring, who essentially develops the "Ramsey rule" for many consumers by bringing redistribution directly into the welfare-maximization problem. Martin Feldstein and H.A. John Green also derive the Ramsey rule with many consumers; they impose a distributional assumption and an assumption about Engel curves, discussed below in fn. 2. Diamond and Mirrlees present optimal tax formulae for nonidentical consumers in their joint paper, and both have taken up the many-consumer problem again in separate papers (Mirrlees, 1975, and Diamond). Their approach differs from that taken here in that they do not analyze the reasons why the many-consumer formulae diverge from those encountered in the identical-consumer case. The spirit of the Mirrlees paper is in some respects similar to that of the present discussion, however. (See fn. 7 below.) Finally, a number of very recent articles have expressed concern with the usual restriction to the single-consumer case, and have emphasized that distributional issues must somehow be dealt with in the study of commodity taxation. See Richard Musgrave, Sandmo (1976), and Atkinson and Stiglitz (1976).

derivative of the welfare function with respect to the household's consumption of that good) is the same for all consumers. If one imposes this restriction on the welfare function, the first of the tax rules mentioned above will hold in the many-consumer case (Proposition 1), but the second will not (Proposition 2). The second neutrality assumption (extended neutrality) that might be imposed is that the marginal social utility of *income* (the change in social welfare associated with a one dollar increase in a household's before-tax income) is the same for all consumers. Under the assumption of extended neutrality, the first tax rule mentioned above is not optimal in the many-consumer case (Proposition 3), while the Ramsey rule is (Proposition 4).

Since these neutrality assumptions are assumptions about the specific welfare function being used, one cannot legitimately *assume* that both of these restrictions simultaneously hold. However, if one invokes the simple neutrality assumption, it turns out that a knowledge of the (potentially empirically discoverable) income derivatives of the household demand functions would be sufficient information to compute a lump sum redistribution of pretax income which, if actually performed, would lead to the satisfaction of the extended neutrality assumption. Should this be done, then obviously both tax rules stated above will be optimal (Proposition 5). If this redistribution is not actually carried out, however, both tax rules cannot simultaneously hold. This shows, as already suggested, that the single-consumer case differs fundamentally from the far more relevant many-consumer situation.

The paper is organized as follows. The first section introduces the model and presents a general many-person tax rule. In Section II, I first derive the two tax rules for the single-consumer case. Next the principal results on the extension of the two rules to the many-person economy are presented, along with some comments on the methodology of this and other studies of the optimal taxation problem. Some brief concluding remarks are found in Section III.

I. The Model and the Optimal Taxation Problem

The framework for analysis of optimal tax problems is now quite familiar, so the model can be presented very briefly. In the economy discussed here there are H households ($h = 1, \ldots, H$), a private production sector (distinguished by the superscript f), and a public production sector (superscript g). The government carries out productive activities because it is responsible for public good provision. Like other productive units, it purchases its inputs; this creates the need for taxation. Here we shall neglect the problem of optimal public good provision, and assume simply that the government has exogenously fixed private good inputs, measured negatively in the net *output* vector $y^g \equiv (y_i^g)$, where $-y_i^g$ is the amount used of the ith private good ($i = 0, 1, \ldots, n$).

The private production sector is assumed perfectly competitive and operates subject to the overall technology constraint

$$\sum_{i=0}^{n} \phi_i^f y_i^f \leq 0 \tag{1}$$

where every component of $\phi^f \equiv (\phi_i^f)$ is constant and strictly positive, and $y^f \equiv (y_i^f)$ is the aggregate *net* output vector (with $y_i^f < 0$ for a factor i). Maximized profits will thus be zero in equilibrium. Note for future reference that if $p \equiv (p_i)$ is the price vector facing producers, profit maximization implies that

$$\frac{p_i}{p_k} = \frac{\phi_i^f}{\phi_k^f} \qquad i, k = 0, 1, \ldots, n \tag{2}$$

Note that these relative prices will be invariant due to the constancy of ϕ^f.

Let $x^h \equiv (x_i^h)$ denote the net consumption vector for household h, where $x_i^h \gtrless 0$ as i is a commodity demanded or supplied. Preferences are represented by strictly quasi-concave differentiable utility functions $u^h(x^h)$. Let $q \equiv (q_i)$ denote the vector of consumer prices; we have $q = p + t$ where $t \equiv (t_i)$ is the vector of commodity taxes. Households select consumption

vectors which maximize utility subject to budget constraints

$$(3) \qquad qx^h = 0$$

The absence of a term on the right-hand side of (3) reflects the absence of lump sum taxes and subsidies, and of private sector profits. Utility maximization yields the household's demand vector as a function of prices and fixed income I^h (which, as just noted, is zero). Let

$$v^h(q,I^h) \equiv u^h(x^h[q,I^h])$$

be the indirect utility function for h. It satisfies Roy's formula,

$$(4) \qquad \frac{v_k^h}{v_I^h} = -x_k^h$$

where $v_k^h \equiv \partial v^h/\partial q_k$, and $v_I^h \equiv \partial v^h/\partial I^h$, the marginal utility of income.

The government attempts to maximize the social welfare function $W(v^1, \ldots, v^H)$ by a proper choice of consumer prices and producer prices. (Optimal taxes are thus determined implicitly.) In doing this it is constrained to obtain the necessary inputs for public good provision y^g, subject to the condition that markets clear. Let $X_i \equiv \sum_h x_i^h$ be the market demand for good i, with $X \equiv (X_i)$. Then market clearing requires that

$$(5) \quad X_i = y_i^f + y_i^g \qquad i = 0,1,\ldots,n$$

The government's budget must also be balanced. Its revenues are tX, while its expenditures are $-py^g$. Hence budget balance requires that

$$(6) \qquad tX + py^g = 0$$

But if all households satisfy their budget constraints, if maximized profits are zero, and if all markets clear, (6) will automatically be satisfied. Hence (6) is a redundant constraint. Note also that if p and q are vectors of producer and consumer prices that satisfy all of the constraints, so are αp and βq for any positive α and β, *not necessarily equal.* Therefore it is possible to normalize *both* price systems; let us suppose that this is done to achieve $q_0 = p_0 = 1$. This implies, of course, that $t_0 = 0$.

Finally, rather than have the government actually choose producer prices, it is formally convenient to have it select a supply vector for the private sector directly; the producer prices that could be used to support this output vector will then be given by the relations (2). It then becomes necessary, however, to append the private production function (1) as a constraint.

The Lagrangian can now be formulated:

$$L = W(\{v^h\}) + \sum_{i=0}^{n} \rho_i(y_i^f + y_i^g - X_i) - \lambda^f \phi^f y^f$$

The first-order conditions are that

$$(7a) \quad \frac{\partial L}{\partial q_k} = \sum_h W_h v_k^h - \sum_{i=0}^{n} \rho_i \frac{\partial X_i}{\partial q_k} = 0$$

$$(7b) \quad \frac{\partial L}{\partial y_k^f} = \rho_k - \lambda^f \phi_k^f = 0 \qquad k = 0,1,\ldots,n$$

where $W_h = \partial W/\partial v^h$, of course.

From (7b) and (2), $\rho_i/\rho_0 = p_i$; substitution from (4) into (7a) thus yields

$$-\sum_h \frac{W_h v_I^h}{\rho_0} x_k^h = p \frac{\partial X}{\partial q_k}$$

Differentiating the sum of all the budget constraints (3) with respect to q_k and substituting, we have

$$(8) \qquad \sum_h \frac{W_h v_I^h}{\rho_0} x_k^h = t \frac{\partial X}{\partial q_k} + X_k$$

With producer prices constant (from (2)), $dt_k = dq_k$, so that (8) becomes

$$(9) \quad \sum_h \frac{W_h v_I^h}{\rho_0} x_k^h = t \frac{\partial X}{\partial t_k} + X_k = \frac{\partial (tX)}{\partial t_k}$$

Equation (9) will serve as a starting point for further derivations.

II. Sufficient Conditions for Two Optimal Tax Rules

While (9) is a general many-person tax rule, it is not expressible as one of the frequently encountered formulae of the single-consumer model, as it stands. Let me

now state formally the two common tax rules alluded to earlier:

Optimal Tax Rule 1 (Marginal revenue proportionality): For every commodity, the ratio of the marginal tax revenues associated with a marginal increase in the tax rate on the commodity to the quantity of the commodity purchased is the same; that is,

$$\frac{\partial(tX)/\partial t_i}{X_i} = \frac{\partial(tX)/\partial t_k}{X_k} \qquad i,k = 1,\ldots,n$$

Optimal Tax Rule 2 (Ramsey rule): For all commodities, the percentage reduction in demand (along the compensated demand curve) due to taxation is (approximately) the same.

A. *Results for a Single Consumer*

Before going on to discuss these rules in a many-person context, it may be helpful to present a brief derivation for the single-consumer case. If there is only one consumer, (9) is simplified by eliminating the social welfare function and by identifying individual and market quantities:

$$\frac{v_I}{p_0} X_k = \frac{\partial(tX)}{\partial t_k} \tag{10}$$

Clearly, division by X_k yields the marginal revenue proportionality rule.

Next, using the Slutsky relation for the derivative of a demand function with respect to a tax rate (= derivative with respect to a consumer price, producer price constant), namely,

$$\frac{\partial X_i}{\partial t_k} = \left.\frac{\partial X_i}{\partial t_k}\right|_{\bar{u}} - X_k \frac{\partial X_i}{\partial I} \tag{11}$$

equation (10) yields

$$\frac{v_I}{p_0} X_k = X_k + \sum_{i=0}^{n} t_i \left.\frac{\partial X_i}{\partial t_k}\right|_{\bar{u}} - X_k t \frac{\partial X}{\partial I}$$

Using the symmetry of the substitution terms (i.e., $(\partial X_i/\partial t_k)_{\bar{u}} = (\partial X_k/\partial t_i)_{\bar{u}}$) and rearranging, we have

$$\frac{v_I}{p_0} - 1 + t\frac{\partial X}{\partial I} = \sum_{i=0}^{n} \frac{t_i}{X_k} \left.\frac{\partial X_k}{\partial t_i}\right|_{\bar{u}}$$

The right-hand side of this expression is the (approximate) percentage reduction in (compensated) demand for commodity k due to taxation. Since the left-hand side is independent of k, we see that this percentage reduction is the same for all commodities: the Ramsey rule. Thus both tax rules are valid and equivalent characterizations of the optimal tax structure for a single-consumer economy.

B. *The Many-Person Case: Simple Neutrality*

From one perspective, the rest of this paper is devoted to the task of finding conditions under which the above tax rules obtain in a many-consumer economy. These conditions will take the form of restrictions on the social welfare function W, which (roughly speaking) state that the distribution of welfare is in some sense equitable. As in so many other cases, there is a tremendous gain in the sharpness of analytical results when one assumes away the distributional problem and focuses (insofar as possible, and not necessarily without ambiguity) on the efficiency issues in isolation. I shall discuss the legitimacy of this approach later.

One interesting assumption that will be used below states that the marginal social utility of consumption of the numeraire good is the same for every household. I shall call this the "simple assumption of distributional neutrality."

ASSUMPTION 1 (simple neutrality): *$W_h u_0^h$ is the same for all h; that is,*

$$W_h u_0^h \equiv \mu \qquad \text{for } h = 1,\ldots,H$$

Note that

$$v_I^h = \sum_{i=0}^{n} u_i^h \frac{\partial x_i^h}{\partial I^h} = u_0^h \sum_{i=0}^{n} q_i \frac{\partial x_i^h}{\partial I^h} = u_0^h$$

It will be more convenient to use Assumption 1 in the marginal utility of income form

$$W_h v_I^h \equiv \mu \tag{12}$$

This is one possible characterization of

neutrality, but it is not the only one, as we shall see below. For the moment, however, let us restrict our attention to Assumption 1, using it to demonstrate two propositions: first, that the simple neutrality assumption is sufficient to establish the optimality of the marginal revenue proportionality rule; second, that simple neutrality is *not* sufficient to establish the Ramsey rule. The fact that this second rule does not generally hold under simple neutrality motivates the search for another definition of neutrality.

Beginning with the general optimal tax rule (9), it is a short step to the marginal revenue proportionality rule under Assumption 1, for substituting (12) into (9), minor rearrangement yields

$$\frac{\partial(tX)/\partial t_k}{X_k} = \frac{\mu}{\rho_0} \qquad k = 1, \ldots, n$$

But this is precisely Optimal Tax Rule 1.

PROPOSITION 1: *Under the simple assumption of neutrality, the marginal revenue proportionality rule is optimal in the many-consumer case.*

On the other hand, using the Slutsky equation (11) and the neutrality condition (12) in (9) produces

$$(13) \quad \frac{\mu}{\rho_0} X_k = \sum_{i=0}^{n} t_i \left.\frac{\partial X_i}{\partial t_k}\right|_{\bar{u}} - \sum_{i=0}^{n} t_i \left(\sum_h x_k^h \frac{\partial x_i^h}{\partial I^h}\right) + X_k$$

where $\partial X_i/\partial t_k |_{\bar{u}}$ is the sum of the price derivatives of the compensated demand functions.

Rearranging (13), using the symmetry of the substitution terms, one has

$$(14) \quad \sum_{i=0}^{n} \frac{t_i}{X_k} \left.\frac{\partial X_k}{\partial t_i}\right|_{\bar{u}} = \frac{\mu}{\rho_0} - 1 + \sum_h \frac{x_k^h}{X_k} \frac{\partial tx^h}{\partial I^h}$$

The left-hand side of (14) is the percentage reduction in demand along the compensated demand for good k. Observe that the right-hand side of (14) will vary from commodity to commodity (i.e., will vary with k) because the share of commodity k consumed by h, x_k^h/X_k, will vary with k; hence, the weights attached to the terms $\partial tx^h/\partial I^h$ will vary. Moreover, if income derivatives of demand vary from consumer to consumer, as one would expect, these latter terms will all be different as well. Hence, only in the special case where each household consumes the same fraction of the aggregate amount of each commodity and/or has identical income derivatives of demand, as would be the case with a single consumer or many identical consumers, will the Ramsey rule be valid.[2]

PROPOSITION 2: *Under Assumption 1, the Ramsey rule is not generally valid in the many-consumer case.*

C. *An Alternative Neutrality Concept and the Methodology of Optimal Tax Analysis*

As mentioned above, Assumption 1 does not imply that the welfare maximizer would not, if possible, carry out lump sum redistribution among households. To see this, suppose that although all *net* government expenditures (i.e., expenditures for public good provision) must be commodity tax financed, so that the balanced budget constraint (6) continues to hold, the government is nevertheless permitted to distribute

[2]Feldstein and Green arrive at the Ramsey rule by in effect imposing Assumption 1 (since the Feldstein-Green "distributional characteristic" is the same for every commodity in this case), and by further imposing the Gorman-Nataf condition that the Engel curves for all individuals are parallel straight lines. To see this result, note that the latter condition implies that for each i, $q_i(\partial x_i^h/\partial I^h)$ is the same for all h. But then $(q_i - p_i)(\partial x_i^h/\partial I^h) = t_i(\partial x_i^h/\partial I^h) = \sigma_i$ is the same for all h. Substituting in $\Sigma_i \sigma_i$ for the last term on the right-hand side of (14) produces an expression independent of k, the Ramsey rule. I am concerned here with the *general* many-person problem, however, and do not wish to impose such restrictions on consumer behavior.

to each household a lump sum amount of the numeraire commodity, s_0^h; this will be negative for some households (corresponding to a lump sum tax). It is required that the *sum* of all these be zero, that is,

$$\sum_h s_0^h = 0 \tag{15}$$

Thus one can imagine that the government is allowed to determine the initial distribution of income via costless lump sum transfers; once this is done, it is restricted to use commodity taxation to achieve a net transfer of resources from the private to the public sector. In this situation, the household's budget constraint would become

$$qx^h = s_0^h \tag{16}$$

Note that $\partial v^h/\partial s_0^h \equiv v_I^h$, $\partial x_i^h/\partial s_0^h \equiv \partial x_i^h/\partial I^h$, etc.

In setting up the new welfare-maximization problem, it is necessary to add (15) as a constraint. The Lagrangian is therefore

$$L = W(\{v^h\}) + \sum_{i=0}^{n} \rho_i (y_i^f + y_i^g - X_i) - \lambda^f \phi^f y^f + \theta \sum_h s_0^h$$

The first-order conditions are as before (equations (7)) with the addition of

$$W_h v_I^h - \sum_{i=0}^{n} \rho_i \frac{\partial x_i^h}{\partial I^h} + \theta = 0 \qquad h = 1, \ldots, H \tag{17}$$

It would seem reasonable to interpret (17) as an alternative formulation of the idea of "equal social marginal utilities of income"; or perhaps the phrase "equal social marginal net benefits of income" would be better. Equation (17) says that one equates the social marginal utility or benefit of income for all households, where this benefit is the marginal welfare of the marginal utility of income (or consumption) to the household *less* the marginal welfare cost of the additional consumption that the household would carry out as the result of an addition to income. Alternatively, differentiating (16) with respect to s_0^h, one can rewrite (17) as

$$\frac{W_h v_I^h}{\rho_0} + \sum_{i=0}^{n} (q_i - p_i) \frac{\partial x_i^h}{\partial I^h} + \theta - 1 = 0$$

or

$$\frac{W_h v_I^h}{\rho_0} + \frac{\partial (tx^h)}{\partial I^h} = 1 - \theta \tag{18}$$

Equation (18) suggests the interpretation that if possible the government would (*ceteris paribus*) redistribute income in favor of those whose additional spending from another dollar of income would generate the most additional taxes.

Obviously equation (18) will typically *not* be satisfied if the government is not permitted to carry out this ideal redistributive scheme. Therefore, to assume that (18) is satisfied in the absence of such a scheme is an assumption about the welfare function W. This assumption has a claim to be considered as a formalization of the notion of distributional neutrality: if met, the welfare maximizer would *choose* not to carry out any lump sum redistribution of income, even if empowered to do so. On the other hand, we note that equation (18) is not equivalent to Assumption 1, simple neutrality. Clearly, $W_h v_I^h = \mu$ for all h neither implies, nor is implied by, equation (18). Thus, even if one assumes simple neutrality, the government would still wish to engage in lump sum redistribution, if it could. This suggests that (12) may not be the best (and certainly is not the only) definition of distributional neutrality. In attempting to isolate and suppress the purely distributional aspects of the problem of optimal commodity taxation, then, one might wish to invoke the following assumption:

ASSUMPTION 2 (extended neutrality): *Even if empowered to do so, the government will not carry out lump sum redistribution; that is, the welfare function W is such as to satisfy equation* (18).

Before going on to indicate the application of Assumption 2, I would like to discuss its relationship to Assumption 1, and

to raise some troubling methodological points.

First, having motivated Assumption 2 by noting that it would be satisfied as a consequence of lump sum redistributional measures, it is useful to reconsider Assumption 1 from the same perspective. Suppose that households possess given consumption bundles, and that the government is able to redistribute the numeraire good in lump sum fashion subject to the condition that no household is permitted to trade away from its posttransfer consumption bundle. Then clearly Assumption 1 will characterize the optimal set of transfers. Thus, one can think of Assumption 1 as a neutrality assumption corresponding to the situation after all trades have taken place and markets are closed, whereas Assumption 2 characterizes distributional neutrality in an *ex ante* sense, allowing for the effects of redistribution on the purchases of households.[3]

Note that in the absence of commodity taxes, all t_i's are zero, (18) reduces to (12), and the distinction between simple and extended neutrality disappears. The possibility of two different neutrality concepts depends essentially on the existence of commodity taxes.

One immediate objection to either of these assumptions is that it is strange to suppose that the government can redistribute income in lump sum fashion but cannot raise *net* revenues in a nondistortionary way. It would seem that if the government could devise optimal lump sum redistributions of income, it should be able to avoid distortionary taxation altogether. On the other hand, one might turn this argument around and observe that if the government has the ability to compute optimal commodity taxes— if it can impose a tax system satisfying condition (9), for example—then it should be able to compute optimal lump sum taxes. For to raise a given amount of revenue in lump sum fashion, the government needs "only" to know the marginal social utility of income or consumption (the two are the same with lump sum taxes) for every consumer, which is clearly less information than is needed for optimal commodity taxes (in the general case). Once given a completely specified social welfare function which can be used to resolve all issues of ethical deservingness, it *is* possible to devise ideal taxes: if it is desired that Paul should be made better off at the expense of Peter, then tax Peter relatively heavily because he is Peter and tax Paul relatively lightly because he is Paul. Such taxes are unrelated to the economic behavior of the individuals and hence lump sum.

I would conjecture that the discussion of optimal commodity taxation has been motivated by the observation that most widely practiced forms of taxation (a) are based on the economic behavior of the individual (for example, income taxes) and (b) seem to be designed, at least in part, to achieve distributional objectives (for example, many income tax systems are progressive; income supplements of various kinds are usually directed toward the poor, etc.). The latter point suggests that somewhere the ethical problems have all been worked out, so that a social welfare function exists; lump sum taxes would certainly be used to maximize this function if feasible; (a) suggests that they are not used; hence, lump sum taxes are infeasible.

But why should lump sum taxes be infeasible? Certainly the physical act of taxation on a lump sum basis is feasible; on the administrative level, lump sum taxes would seem to involve smaller collection costs than a complex system of commodity taxation.[4] And as just noted, the informational requirements for optimal lump sum taxation are less severe than those for optimal commodity taxes. I think it should be clear that there is nothing infeasible about optimal lump sum taxes except the construction of the social welfare function needed to

[3] I should note here that Diamond also distinguishes between the marginal social utilities of income and consumption, which underlies the distinction between simple and extended neutrality; he does not, however, introduce any neutrality assumption.

[4] Walter Perrin Heller and Karl Shell have discussed the costs of commodity taxation and their implications for production efficiency.

compute them; but then optimal *commodity* taxes may also be infeasible.

This discussion obviously raises serious questions about the whole approach to the study of optimal commodity taxation as an exercise in social welfare maximization.[5] On the other hand, setting up a welfare-maximization problem does permit one to characterize the set of Pareto (quasi-) optimal tax systems,[6] and it may be possible to find particular systems which can be implemented without specific reference to an underlying social welfare function. This offers hope that the welfare-maximization technique can be used to obtain relatively practicable tax rules which at least permit the attainment of Pareto optimality.[7]

I would hasten to point out however that tax rules which do not explicitly involve the social welfare function have at best tenuous claims to special attention as ethically neutral "efficient" tax rules. Each of the two neutrality assumptions introduced above leads to a different tax formula, as reference to Propositions 1 and 4 (below) will verify; which formula, if either, describes the efficient tax system? Each rule leads to a particular point on the Pareto frontier, and there is no efficiency basis for choosing between them. There is simply no unambiguous way to segregate the efficiency and distributional aspects of the optimal tax problem; however much one is tempted to apply the intuitive insights derived from the study of the single-person economy, they are of limited value in the more interesting many-person context.

The foregoing discussion has raised some perplexing questions about the proper approach to the study of optimal taxation problems and about the interpretation of certain results. Unfortunately I shall not be able to satisfactorily resolve these questions here, and in the remainder of this paper, I shall proceed uncritically within the standard framework. Assumptions 1 and 2 have a number of implications which shed some light on the critical differences between single-consumer and many-consumer economies; these should be of interest in themselves, however one feels about their interpretation in view of the above remarks.

D. *Extended Neutrality and Optimal Taxation*

Some of the implications of the simple neutrality assumption have already been examined and summarized in Propositions 1 and 2. Let us now exploit the extended neutrality assumption in a similar fashion.

Substituting from equation (18) into (9), and dividing through by X_k yields

$$(19)\quad (1-\theta) - \sum_h \frac{x_k^h}{X_k}\,\frac{\partial(tx^h)}{\partial I^h} = \frac{\partial(tX)/\partial t_k}{X_k}$$

which does not reduce to the marginal revenue proportionality rule because the weight x_k^h/X_k that is attached to each term in the sum on the left-hand side will depend on k. Only if every household purchases the same fraction of each commodity or has the same income derivatives of demand for all commodities, as would be true with one consumer or many identical consumers, will the extended neutrality assumption be sufficient to establish Optimal Tax Rule 1.[8] This demonstrates

PROPOSITION 3: *Under Assumption 2, the*

[5]Hahn's comments, pp. 105–06, are particularly to the point here.

[6]Since $W_h' > 0$ for all h, any tax system satisfying the general condition (9) must be Pareto optimal. (These are second best optima, of course.) The selection of a particular W function singles out one of the points along the Pareto frontier as best; as we vary the form of the welfare function, we trace out what Paul Samuelson would call the utility-feasibility frontier.

[7]As Hahn has said, "Optimum tax formulas are either guides to action or nothing at all" (p. 106). Will tax rules that depend on the partial derivatives of a social welfare function ever be suitable guides to action? In this regard compare the remarks of Mirrlees in discussing his two-class rule: "The particular appeal of the result is that it does not refer to the relative social marginal utilities of the two classes" (1975, p. 30). Diamond's rule—"the change in aggregate compensated quantity demanded is proportional to the covariance between individual quantities demanded and social marginal utilities of income" (p. 335)—though perhaps simpler in form than other statements of the general many-person rule, does not appear to lend itself to ready application.

[8]See (14) above. Note that the Gorman-Nataf condition of fn. 2 will again do the trick; making the appropriate substitution in the left-hand side of (19) gives an expression independent of k.

marginal revenue proportionality rule is not valid in the general many-person case.

Next, using the Slutsky equation in the right-hand side of (19), we have

$$1 - \theta - \sum_h \frac{x_k^h}{X_k} \frac{\partial(tx^h)}{\partial I^h} = 1 + \sum_{i=0}^{n} \frac{t_i}{X_k} \frac{\partial X_i}{\partial t_k}\bigg|_{\bar{u}} - \sum_h \frac{x_k^h}{X_k} \frac{\partial(tx^h)}{\partial I^h}$$

By virtue of the symmetry of the substitution terms,

$$\sum_{i=0}^{n} \frac{t_i}{X_k} \frac{\partial X_k}{\partial t_i}\bigg|_{\bar{u}} = -\theta \qquad (20)$$

Thus we have demonstrated

PROPOSITION 4: *Under Assumption 2, the Ramsey rule is optimal in the many-person case.*

To digress momentarily, it is interesting to note from (20) that the percentage reduction in demand along the compensated demand curve should equal the negative of the Lagrange multiplier for the constraint (15). In general, a Lagrange multiplier shows the marginal effect on the objective function of a slight easing of the associated constraint;[9] in the present case, this means that θ is the marginal welfare associated with a one dollar increase (from zero) of the amount of *net* revenues raised via lump sum taxes. It is the "distortion" of commodity taxes that causes the divergence from the first best situation, and so it is not surprising to find a relationship between changes in demand and the welfare cost of commodity taxes; but, intuitively, it is not clear that one would expect to find this precise relationship. In any case, the result would seem to be sufficiently striking to merit special mention.[10]

PROPOSITION 4': *Under Assumption 2, the percentage reduction in demand along the compensated demand curves is equal to minus the welfare gain associated with a marginal increase in the amount of net revenues that may be raised via lump sum taxes.*

Finally, suppose that one invokes both Assumptions 1 and 2. Of course, both of these cannot hold as *assumptions* about the welfare function, since each would specify a *different* function. Rather, one can think of assuming simple neutrality and then (in line with our earlier remarks) actually carrying out a pretax redistribution of income so as to achieve the satisfaction of (18). In this situation, the following result can easily be established:

PROPOSITION 5: *When both Assumptions 1 and 2 hold, both the marginal revenue proportionality and Ramsey rules are optimal (and equivalent) in the general many-consumer case.*

III. Conclusion

The major objective of this paper has been to analyze the problem of optimal taxation in a many-consumer economy, to see whether and under what circumstances certain familiar results from the single-consumer case apply in this more interesting context. One might expect, perhaps, that the set of such circumstances is empty due to the inherent complexities introduced by nonhomogeneous tastes. Alternatively, one might think that the restriction to a single consumer has been imposed in an attempt to analyze the efficiency aspects of the optimal tax problem, leaving aside the equity or distributional issues. As we have seen, the latter view is more nearly the correct one: under certain purely distributional assumptions, the standard one-consumer results do apply in the many-person case. To be more precise, both the marginal revenue proportionality and Ramsey rules will result in Pareto (quasi-) optimal allocations, given that commodity taxes must be used for raising revenues. However, there is no efficiency criterion for choosing between these two rules, nor is there any reason to believe that either will result in a particularly desirable

[9]See, for example, Michael Intriligator, pp. 36–38, 60–62.

[10]Atkinson and Stern recognize that θ is the marginal welfare gain associated with an increase in net lump sum revenues (see their Appendix, pp. 126–27). However, they do not appear to notice the connection with the Ramsey rule.

allocation (defining "desirable" in terms of some ethical criterion).

Indeed, distributional questions are really of the essence of the commodity taxation problem. After all, if one were merely concerned to achieve a Pareto optimum, one could dispense with commodity taxes altogether: a uniform head tax ensures a distortionless *first* best optimum. Thus the real relevance of the frequently cited propositions of the efficiency-oriented single-person model remains obscure. Moreover, as my earlier remarks indicate, the incorporation of the distributional aspects of the problem in the familiar social welfare function may also leave no natural motivation for the introduction of commodity taxes. Perhaps the whole approach to the study of optimal commodity taxation as an exercise in social welfare maximization of the usual type is due for reconsideration.

REFERENCES

A. B. Atkinson and N. H. Stern, "Pigou, Taxation, and Public Goods," *Rev. Econ. Stud.*, Jan. 1974, *41*, 119-28.

——— **and J. E. Stiglitz,** "The Structure of Indirect Taxation and Economic Efficiency," *J. Publ. Econ.*, Apr. 1972, *1*, 97-119.

——— **and** ———, "The Design of Tax Structure: Direct Versus Indirect Taxation," *J. Publ. Econ.*, July/Aug. 1976, *6*, 55-75.

W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, *60*, 265-83.

M. Boiteux, "Sur la gestion des Monopoles Publics astreints à l'équilibre budgétaire," *Econometrica*, Jan. 1956, *24*, 22-40; translated by W. J. Baumol and D. F. Bradford as "On the Management of Public Monopolies Subject to Budgetary Constraints," *J. Econ. Theory*, Sept. 1971, *3*, 219-40.

P. Dasgupta and J. E. Stiglitz, "On Optimal Taxation and Public Production," *Rev. Econ. Stud.*, Jan. 1972, *39*, 97-103.

P. Diamond, "A Many Person Ramsey Tax Rule," *J. Publ. Econ.*, Nov. 1975, *4*, 335-42.

——— **and J. A. Mirrlees,** "Optimal Taxation and Public Production," *Amer. Econ. Rev.*, Part I, Mar. 1971, *61*, 8-27; Part II, June 1971, *61*, 261-78.

A. Dixit, "On the Optimum Structure of Commodity Taxes," *Amer. Econ. Rev.*, June 1970, *60*, 295-301.

J. Drèze, "Some Postwar Contributions of French Economists to Theory and Public Policy, with Special Emphasis on Problems of Resource Allocation," *Amer. Econ. Rev.*, June 1964, Suppl., *54*, 1-64.

M. S. Feldstein, "Distributional Equity and the Optimal Structure of Public Prices," *Amer. Econ. Rev.*, Mar. 1972, *62*, 32-36; "Errata," Sept. 1972, *62*, 763.

H. A. J. Green, "Two Models of Optimal Pricing and Taxation," *Oxford Econ. Pap.*, Nov. 1975, *27*, 352-82.

F. H. Hahn, "On Optimum Taxation," *J. Econ. Theory*, Feb. 1973, *5*, 96-106.

W. P. Heller and K. Shell, "On Optimal Taxation with Costly Administration," *Amer Econ. Rev. Proc.*, May 1974, *64*, 338-45.

Michael D. Intriligator, *Mathematical Optimization and Economic Theory*, Englewood Cliffs 1971.

J. A. Mirrlees, "On Producer Taxation," *Rev. Econ. Stud.*, Jan. 1972, *39*, 105-11.

———, "Optimal Taxation in a Two-Class Economy," *J. Publ. Econ.*, Feb. 1975, *4*, 27-33.

H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, *60*, 693-705.

R. A. Musgrave, "ET, OT and SBT," *J. Publ. Econ.*, July/Aug. 1976, *6*, 3-16.

F. P. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, *37*, 47-61.

P. A. Samuelson, "Evaluation of Real National Income," *Oxford Econ. Pap.*, Jan. 1950, *2*, 1-29.

A. Sandmo, "A Note on the Structure of Optimal Taxation," *Amer. Econ. Rev.*, Sept. 1974, *64*, 701-6.

———, "Optimal Taxation—An Introduction to the Literature," *J. Publ. Econ.*, July/Aug. 1976, *6*, 37-54.

J. E. Stiglitz and P. Dasgupta, "Differential Taxation, Public Goods, and Economic Efficiency," *Rev. Econ. Stud.*, Apr. 1971, *38*, 151-74.

Involuntary Saving through Unanticipated Inflation

By ANGUS DEATON*

In a situation where the consumer, every time he makes a purchase, knows accurately his income and the prices of all the goods which make up his budget, there is no reason to suppose that the pattern of his purchases will be disturbed by a proportional change in these variables. In reality, however, purchases are made sequentially, some frequently and some infrequently, so that while the purchaser has exact up-to-date information about the prices of the goods actually being bought, he does not have accurate knowledge of the prices of other goods. For example, on a shopping trip for beef, the price of beef will be observed, as will almost certainly the prices of other meats; small additional effort will yield the prices of other foods, especially in a supermarket, but it is likely to be prohibitively expensive, not to say troublesome, to observe the prices of automobiles, cameras, take-away Chinese meals, foreign holidays, and embroidered waistcoats. Nor can this information, on which the consumer has only outdated observations, be effectively replaced by knowledge of a general price index. Official indices of retail prices are published with a considerable lag and are not relevant to any given consumer whose income, family size, and tastes will differ from those assumed by the index. On the other hand, compilation of a private index is likely to require at least one full cycle of purchases, again involving a considerable passage of time.

In consequence, at least in the first instance, *individual consumers have no possible means of distinguishing relative price changes from absolute price changes.* For example, in a time of accelerating inflation, assume that a consumer expects prices to have risen by 1 percent since his last shopping expedition whereas the true figure is 2 percent. On entering the grocery store he discovers that coffee, the relative price of which has not in fact changed, is 1 percent more expensive than expected; he may then buy less coffee. If he is more sophisticated (and has enough time), he will inspect prices of other foodstuffs, and provided he can satisfactorily allow for genuine *relative* price changes, he will discover after some calculation that food is more expensive than he had expected; he may then buy less food, not merely less coffee. Eventually, of course, he will discover his mistake, but only after the full range of relative prices has been observed. A similar situation faces other consumers so that, at any given moment of time, different individuals are engaged in the purchase of different commodities, and each finds that the price of that commodity is higher than expected, while none at that instant has enough information to calculate the absolute price level. Consequently, there is a mass illusion that all goods are relatively more expensive so that, as each consumer attempts to adjust his purchases, real consumption falls, and if real income is maintained, the saving ratio rises. Certainly, as mistakes are discovered, attempts will be made to rectify them, but if inflation continues to accelerate and if expectations lag behind reality, the saving ratio will remain abnormally high. By a similar argument, a decelerating rate of inflation will be associated with an abnormally low saving ratio.

It is the purpose of this paper to present theoretical and empirical results which formalize and support this impressionistic description. The model which is discussed below is a *disequilibrium* model of behavior; it is consistent with any formulation of the

*University of Bristol. A large number of individuals made important and helpful comments on earlier drafts of this paper. I should like particularly to thank John Flemming, Charles Goodhart, Frank Hahn, Robert Neild, David Newbery, Peter Phillips, John Muellbauer, Richard Stone, Henri Theil, and Roger Witcomb.

equilibrium consumption function and, in the theoretical development, no particular assumptions will be made about the latter. Instead, the focus is on the means by which the absence of complete price information causes mistakes to be made. The discussion begins from a single consumer buying a single commodity. In ideal conditions, these individual purchases will add up to desired total consumption, as predicted by the equilibrium consumption function. But since no one consumer knows all prices, or even a current index of all prices, mistakes are made, so that the aggregation of purchases over different consumers can produce systematic deviations of actual from equilibrium consumption. A useful analogy is provided by the familiar consumption function in which expenditures are a function of expected, or permanent, income. Deviations of actual from expected income induce unanticipated fluctuations in savings and to a lesser extent in expenditures. In the model discussed here, it is recognized that the sequential nature of purchasing, given the way price information is disseminated, means that the expected and actual aggregate price levels can differ not only in the future but, more significantly, in the present. The savings function thus contains three terms, not two: a term derived from the equilibrium consumption function; a term in unanticipated or transient income; *plus* a term in unanticipated prices. Such a savings function is derived formally in Section I below.

In Section II of the paper, an attempt is made to discover whether or not the effects predicted do or do not exist. This is done by means of an econometric analysis of quarterly American and British data over the last twenty years. No attempt is made to present a comprehensive study of the aggregate consumption function for either country and, in particular, a highly simplistic view is taken of the equilibrium consumption function. This rules out many other possible formulations of the way in which inflation effects saving, but this is consistent with our immediate purpose, to discover to what extent the evidence can be explained in terms of the effects proposed in this paper. To summarize briefly: the existence of the inflation effect is unequivocally clear in both countries; indeed there is a good deal of uniformity in estimated magnitudes between the two. The effects appear to be large, with recent inflation rates capable of shifting the saving rates of both countries well outside the ranges observed in previous postwar experience. Significantly, for neither country does the detection of the inflation effect depend on the inclusion of recent data; even before the fluctuations in the rate of inflation which began in the late 1960's, there is clear evidence of a positive relationship between unanticipated inflation and the savings ratio. This confirms that the results do not depend on the secular correlation between high savings ratios and rising prices over the last few years; indeed the association between quarter-to-quarter changes in the savings ratio and in the rate of price change is immediately visible on inspection of the two series.

Section III contains a review of other studies relevant to the relationship between saving and inflation. It will be argued that, since most of the alternative mechanisms which have been proposed derive their effects through changes in the equilibrium consumption function, they are consistent in principle with the effects proposed here. There is no reason to suppose that any phenomenon has only one cause. However, I shall claim that each of these other mechanisms, taken by itself, has difficulty in offering plausible explanations for the evidence Some concession must be made to the disequilibrium phenomena proposed here.

I. A Disequilibrium Model of Demand

Let us denote money income by y, money expenditure on goods and services by c, money savings by s, quantities purchased by the vector q, and the prices of these quantities by the vector p. The consumer is assumed to make his consumption decision in two parts: at the first stage, equation (1), total expenditure is determined; at the second, equation (2), commodity demands are

given as a function of total expenditure and prices. Hence, for some functions h and f_i,

$$c = h(y, z) \tag{1}$$

$$q_i = f_i(p, c) \tag{2}$$

where z is a vector of relevant variables, for example, wealth, anticipated future prices and income, interest rates, and so on. Note that the possibility that c is a function of the anticipated rate of inflation is not precluded; indeed it is likely to be so, if only through the conventional income and substitution effects of changes in the real rate of interest. The two-stage process (1) and (2) is consistent with the intertemporal utility maximization provided the utility functional is separable over time; this does not, of course, rule out a less formal interpretation.

The demand functions (2) have the property

$$\sum_{i=1}^{n} p_i q_i = \sum_{i=1}^{n} p_i f_i(p, c) = c \tag{3}$$

where n is the number of commodities in the budget. Equation (3) may appear to be little more than an identity, indeed in empirical demand analysis it is usually treated as such. However, the essence of the model presented here is that (3) does not hold for the aggregate of consumers. Total expenditure will be the sum of whatever expenditures by individuals on individual commodities happen to be, mistakes and all, and there is no guarantee that the sum of the c's which enter the individual consumer's demand function will be the same as the actual sum of total expenditures. At any given moment, no single consumer knows the complete price vector p, nor is any one consumer buying all the goods in his budget simultaneously. Each purchaser is buying a different commodity so that each has accurate information on a different subset of prices. Consequently, the price at which a given commodity is actually purchased is not necessarily the same price which different consumers take into account when buying other commodities. The prices at which demands become effective are not necessarily those at which consumers expect to be able to fulfill their notional demands. There is thus no reason why (3) should hold in aggregate unless all prices are correctly anticipated.

At instant t, assume a given consumer is purchasing good i. Assume that immediately prior to purchase, he anticipates an income y^*, and that on the basis of y^*, equation (1), and an anticipated price vector p^*, he plans to spend a total of c^*. At the time of purchase, an accurate value for p_i is observed. Assume (a) that the consumer remains on his demand curve, and (b) that, at the instant t, p_j^*, for all j not equal to i, remains unchanged. (This latter assumption is not essential and the effects of relaxing it will be discussed below.)

We may thus write

$$q_i = f_i(p_1^*, \ldots p_{i-1}^*, p_i, p_{i+1}^*, \ldots p_n^*, c^*) \tag{4}$$

or since $p_i = p_i^* + (p_i - p_i^*)$, taking a Taylor expansion

$$q_i \simeq f_i(p^*, c^*) + (p_i - p_i^*) \frac{\partial f_i}{\partial p_i} \tag{5}$$

provided $(p_i - p_i^*)$ is small or f_i is approximately linear. Writing q_i^* for the first term on the right-hand side and multiplying by p_i, we get on rearrangement

$$p_i q_i = p_i^* q_i^* + (p_i - p_i^*)\left(q_i^* + \frac{\partial f_i}{\partial p_i} p_i\right) \tag{6}$$

Aggregation over consumers and commodities, still at the instant t, is performed first by summing over all consumers buying good i to give the total effective demand for that good, and second, by summing over all such groups to give total instantaneous expenditure. The simplest way to do this is to assume all consumers have identical tastes and identical incomes, but this is grossly oversufficient and can be replaced by the natural assumption that the errors of price expectation are distributed over consumers independently of both quantities consumed and elasticities of demand. Equation (6) can then be summed over i to give

aggregate expenditures c, thus

$$(7)\quad c = \sum_k p_k q_k = c^* + \sum_k q_k^*(1 + e_{kk})(p_k - p_k^*)$$

where e_{kk} is the own-price (uncompensated) elasticity of demand for good k, and we have accepted the approximation $e_{kk} = \frac{\partial f_k}{\partial p_k}\frac{p_k}{q_k^*}$.

The original notation for c has been retained for aggregate effective demand over consumers although originally designated for total notional demand for a single consumer. This emphasizes that in the model, aggregation over commodities is accomplished by aggregating over individuals; c and c^* will be used in the aggregate sense from now on.

From (7), defining the value shares $w_k^* = p_k^* q_k^* / c^*$,

$$\log c = \log c^* + \log\left\{1 + \sum_k w_k^* \cdot (1 + e_{kk})(p_k - p_k^*)/p_k^*\right\}$$

thus

$$(8)\quad \log c \simeq \log c^* + \sum_k w_k^* \cdot (1 + e_{kk})(\log p_k - \log p_k^*)$$

It is convenient, although not essential, to express (8) in terms of a single price index P, rather than in terms of the individual prices. Write

$$(9)\quad \log P - \log P^* = \sum_k w_k^* \; (\log p_k - \log p_k^*)$$

to define P as a Divisia chain index. Define v_k by

$$(10)\quad \log p_k - \log p_k^* = (\log P - \log P^*) + v_k$$

so that the unanticipated proportional change in the kth price is ascribed partly to specific factors v_k, and partly to general unanticipated inflation. We may then either study general inflation effects as such by assuming $v_k = 0$, *ex hypothesi*, or, from an empirical point of view, take this as the dominant effect given that $\Sigma w_k^* e_{kk} v_k$ is likely to be small. Either way, we have

$$(11)\quad \log c \simeq \log c^* + (1 + \phi)(\log P - \log P^*)$$

where, for convenience

$$(12)\quad \phi = \Sigma w_k^* e_{kk}$$

Finally, using the approximation for the savings ratio,

$$\frac{s}{y} \simeq -\log \frac{y - (y - c)}{y} = \log y - \log c$$

we may rewrite (11) as

$$(13)\quad \left(\frac{s}{y}\right) = \left(\frac{s^*}{y^*}\right) + \left\{\log\left(\frac{y}{P}\right) - \log\left(\frac{y^*}{P^*}\right)\right\} - \phi\{\log P - \log P^*\}$$

Equation (13) is the main theoretical result of the paper.

In interpreting (13), it is convenient to begin with the quantity ϕ. From (12), provided that demand curves slope downwards, ϕ must be negative. Its absolute magnitude depends on the way in which price information is conveyed and hence on the structure of retailing and shopping habits. The more commodities are disaggregated, the greater the possibilities for substitution, and thus the greater the absolute magnitude of the own-price elasticities. Consequently, ϕ will be absolutely small if goods are purchased together, say in a hypermarket, while ϕ will be large if commodities are bought one by one, each in a different store. It can also be seen that if assumption (b) above is relaxed, the value of ϕ will be modified without further change to the model. For if consumers react to an unanticipated change in one price by modifying their expectations of other prices, the perceived (and mistaken) relative price movement in the good being purchased will be correspondingly reduced.

But, given overall homogeneity of the demand functions, this is precisely equivalent to a reduction in the own-price elasticities, and thus in ϕ. The sign will still be negative provided consumers do not interpret *every* unanticipated price change as a symptom of general inflation or deflation. Given this, the determinants of ϕ, apart from a scaling factor, will be identical to those outlined above.

We can now interpret (13) term by term. The first term relates to the equilibrium consumption function and gives the savings ratio under the assumption that current expectations of prices and real incomes are correct. Note that the mistakes in previous periods will influence this term. For example, if the long-term equilibrium consumption function is of the type proposed by Milton Friedman, that is,

$$c = ky^{p} \tag{14}$$

then, taking account of previous mistakes, we might specify

$$\log c_t^* = \log k + \log \cdot \left[y_t^* + \sigma \int_{-\infty}^{t} \exp\{(t - \theta)r\}(c_\theta^* - c_\theta)d\theta\right] \tag{15}$$

where r is the rate of interest and σ is the reciprocal of the time horizon and measures the rate at which wealth is absorbed into permanent income. Equation (15) and the value of s^*/y^* that it implies are only possibilities; how one might wish to specify the desired savings ratio depends on the view one takes of the aggregate consumption function and of the way in which mistakes are corrected.

The second term of (13) is also familiar; it reflects the fact that all unanticipated income is saved. This is inevitable given the definitions; the consumer cannot react to a stimulus he does not perceive. Over any finite time period however, unanticipated income will affect anticipated income and some of the former will be consumed.

The final term is the one with which we are primarily concerned and which describes the inflation effect on savings. If, for example, real income is correctly anticipated either by indexation *de jure*, or by wage inflation giving indexation *de facto*, unanticipated inflation will cause the savings ratio to rise. If, however, incomes are not indexed, so that unanticipated price rises cause unanticipated cuts in real income, the savings ratio will fall if, as seems to be the case, ϕ lies between zero and -1. But this fall in the savings ratio is more than accounted for by the unanticipated fall in real income; taken alone the unanticipated rise in the price level works in the opposite direction. Consequently, a conventional consumption function relating saving and consumption to real income, and estimated over a period when inflation was more or less correctly anticipated, will underestimate the savings ratio in a period of unanticipated inflation. The last term in (13) is thus responsible for increasing the savings ratio (when $P > P^*$) over the level predicted by the (conventional) first two terms; unanticipated inflation causes involuntary (and unanticipated) saving.

II. Empirical Analysis of American and British Data

This section presents a summary of empirical material relevant to assessing the consistency between the hypothesis and the data. (A fuller set of results are available from the author.)

The data are quarterly seasonally adjusted observations on income, expenditure, and prices from the mid-1950's for the United Kingdom and the United States. The British data are taken from *Economic Trends* and run from 1955-III to 1974-III, while the American data are from *Survey of Current Business*. The available series for the United States go back to 1947 but a sample period of 1954-II to 1974-II was selected in order to avoid the effects of the Korean War. For the United Kingdom the published figures for personal disposable income were used; the comparable data for the United States were computed by deducting interest paid and personal payments to foreigners. For both countries, expenditure on consumer durables was included in

consumption. At this stage in the development of the model, no attempt has been made to allow for any special relationship between purchases of durables and inflation. Since we are attempting to observe a negative association between inflation and consumption, it seemed best, from a methodological point of view, to include durables in consumption since this is the least favorable procedure for the confirmation of the hypothesis.

In order to derive an estimating equation, we must specify an equilibrium consumption function and mechanisms for modeling expectations. For the former, we take the simplest possibility and take the Friedman model (14) to specify the long-run equilibrium relation between c^* and y^*. Allowance is made for past mistakes by writing this in the form

$$(16) \quad \frac{d}{dt}\left(\frac{s^*}{y^*}\right) = \sigma\left\{(1 - k) - \left(\frac{s}{y}\right)\right\}$$

According to this the savings ratio is constant at the equilibrium level $(1 - k)$; otherwise, its rate of change is proportional to the current shortfall. Ideally, it would be preferable to use (15) rather than (16); however (15) presents severe estimation problems, (16) is quite sensible as an alternative, and is in any case a close approximation to the derivative of (15). Substitution of (16) in (13) after differentiation yields

$$(17) \quad \frac{d}{dt}\left(\frac{s}{y}\right) = \sigma(1 - k) + (\rho - \rho^*) - \phi(\pi - \pi^*) - \sigma\left(\frac{s}{y}\right)$$

where ρ and π are the instantaneous rates of growth of real income and the price level. Both k and ϕ will be treated as constants. Note that the constancy of k rules out movements in the desired savings ratio due to changes in the anticipated rate of inflation or other variables. This is not *necessarily* incorrect since the income and substitution effects of lower real interest rates work in opposite directions. However, the main reason for adopting this procedure is not out of realism, but to see how far we can get using disequilibrium effects alone. Alternative explanations will always be possible, and some of these will be considered in the next section. The parametrization of ϕ is more straightforward. It is a dimensionless quantity, and although one might expect it to have a secular downward trend due to changes in retailing structures, there is no reason to suppose that it is not independent of P and y. Although not strictly relevant in aggregate, one can choose preference orderings for a single consumer which guarantee the approximate constancy of (12); the choice of ϕ as a parameter is thus unlikely to introduce any unwelcome implicit assumptions.

The main problem that remains is the specification of the mechanisms linking inflation and real income expectations to their actual values. Again, we begin with a very simple specification and assume that

$$(18) \quad \rho^* = \beta\rho + (1 - \beta)\rho^0; \quad 0 \leq \beta \leq 1$$

$$(19) \quad \pi^* = \pi^0$$

where β, ρ^0, and π^0 are constants. Equation (18) is an adjustment mechanism similar to but much simpler than the Koyck formulation; it postulates that the expected rate of growth of real income is a weighted average of the actual rate of growth and some base, or normal rate of growth ρ^0. Since there has been little apparent trend in the actual values of ρ in either country over the sample, it is not unreasonable to use a constant value for the normal rate of growth. The parameter β should be thought of as dependent on the length of the observation period. As time goes on and the consumer observes the actual changes in prices and income, ρ^* will tend to ρ and β to unity; if the time period is very short, β will be much closer to zero. Equation (19) makes the simplest possible (and clearly unrealistic) assumption about π^*; however it is only unreasonable for the most recent observations and may be a fair approximation for most of the sample. There is no point in generalizing (19) to the form of (18); the extra parameter would not solve the problem of an upward trend in π and in any case would not be separately identifiable.

I have taken the view that real income and inflation are the variables for which expectations are formed, not money income and inflation. This assumption has been made because the greater stability of real income changes should render them more easily predictable; nevertheless, expectations may not be formed in this way, and real income expectations may be derived indirectly from expectations on money income and prices. If this is the case, and if (18) holds for changes in money income while (19) remains true for inflation, equation (20) below will be unchanged, although the interpretation of the constant and the coefficient on the inflation rate will change. This will affect the measurement of ϕ only; the relationship between saving, income, and inflation would not be altered.

Substitution of (18) and (19) in equation (17) gives

$$(20)\quad \frac{d}{dt}\left(\frac{s}{y}\right) = \{\sigma(1-k) - (1-\beta)\rho^0 + \phi\pi^0\} + (1-\beta)\rho - \phi\pi - \sigma\left(\frac{s}{y}\right)$$

Note that our assumptions mean that k, ρ^0, and π^0 are not separately identified.

Estimation of (20) on discrete data requires that the model be integrated over some finite time span [t, t − h], where h may be a quarter, a year, or some other interval. The details are relegated to an Appendix since the calculations closely follow those published by Peter Phillips. The estimating equation corresponding to (20) becomes

$$(21)\quad \left(\frac{s}{y}\right)_t - \left(\frac{s}{y}\right)_{t-h} = (1-e^{-\sigma h})\{(1-k) - \sigma^{-1}(1-\beta)\rho^0 + \sigma^{-1}\phi\pi^0\} + \alpha(1-\beta)\rho_t - \alpha\phi\pi_t - (1-e^{-\sigma h})\left(\frac{s}{y}\right)_{t-h} + u_t$$

where ρ_t and π_t are the actual rates of growth of real income and the price level over the interval [t, t − h], u_t is an error of approximation, and α is a quantity which varies with the interval h according to

$$(22)\quad \alpha = \frac{3}{2\sigma h} - \frac{1}{(\sigma h)^2} - \left(\frac{1}{2\sigma h} - \frac{1}{(\sigma h)^2}\right)e^{-\sigma h}$$

Initially, we take h to be one quarter so that α is fixed and (21) may be treated as a linear regression equation. Estimated on American data, this gives

$$(23)\quad \Delta\left(\frac{s}{y}\right)_t = \underset{(1.86)}{.00830} + \underset{(5.36)}{0.4534}\rho_t + \underset{(4.00)}{0.5536}\pi_t - \underset{(3.57)}{0.2401}\left(\frac{s}{y}\right)_{t-1/4}$$
$$R^2 = 0.3614$$

while on British data, the result is

$$(24)\quad \Delta\left(\frac{s}{y}\right)_t = \underset{(1.39)}{.00616} + \underset{(9.43)}{0.7320}\rho_t + \underset{(3.77)}{0.5936}\pi_t - \underset{(3.62)}{0.2186}\left(\frac{s}{y}\right)_{t-1/4}$$
$$R^2 = 0.5937$$

The figures given in parentheses are t-ratios, while Δ denotes the backwards first difference operator.

For both countries the coefficients of all three variables are highly significant and have the expected signs. Although the coefficients on ρ_t differ between the United States and the United Kingdom—perhaps reflecting the relative unpredictability of changes in real income in the former—the estimates of the adjustment and inflation parameters are very close to one another. Both equations would suggest that when the quarterly rate of inflation is running at 2 percent above the anticipated rate (and this is not absurd on recent experience), slightly more than 1 percent of disposable income is involuntarily saved. Given the historically low variation in the saving ratio, this is a sizeable effect.

Equations (23) and (24) can be refined in various ways. Allowing for first- and fourth-order serial correlation makes very little difference to either equation although there is some evidence of first-order autocorrelation in the residuals for the United King-

dom. More importantly, it is possible to reestimate each equation on the first half and second half of each sample separately. This allows a check on whether the inflation effect has been operative throughout the postwar period or whether it has only existed, or at least been observable, in association with the very high rates of inflation of recent years. In both countries, the inflation variable retains its sign and its significance over both halves of both samples; an inflation effect is clearly detectable in the period from 1955–65.

Attempts to generalize the simple (and in the case of inflation, unrealistic) expectation mechanisms (18) and (19) met with little success. Replacement of ρ^0 and π^0 by declining geometric lags of past values of the respective growth rates did not improve the equations, nor did the use of moving averages. In all cases the rate at which expected inflation adjusted to actual rates was implausibly slow. A less direct way of examining expectations is to use (22) and reestimate the basic equation (21), not only on one-quarter first differences, but also on half-year, three-quarter, and annual differences. As we lengthen the period of observation, and since expectations adapt, the difference between actual and expected changes in real income and inflation can be expected to diminish. Consequently, as we move from quarters to years, the parameter estimates on ρ and π will become smaller. The issue is complicated by the fact that because of changes in the observation period h, the quantity α will change according to (22), thus further modifying the parameter estimates. However, this can be explicitly allowed for since the formula tells us how much the coefficients should change if the model is true and if expectations of changes in real income and inflation are indeed constant.

Table 1 lists the results. In order to preserve as many observations as possible, overlapping differences were used based on the original quarterly data; this has the un-

TABLE 1—CONTINUOUS TIME EXPERIMENTS

Period h	ρ	π	(s/y)	d.w.	ν_1	R^2
			U.S. **Data**			
			No Allowance for Serial Correlation			
1/4	.45336	.55358	−.24013	1.8050		.3614
1/2	.34588	.46553	−.43454	0.8221		.4467
3/4	.26328	.43614	−.63585	0.6846		.5149
1	.20564	.40416	−.80935	0.5917		.5943
			With Allowance for First-Order Serial Correlation			
1/4	48076	.65983	−.31209		.16543	.3446
1/2	.37364	.53666	−.67926		.68637	.6397
3/4	.29019	.44392	−.81362		.69968	.7243
1	.27681	.47001	−1.0009		.75089	.8018
			U.K. **Data**			
			No Allowance for Serial Correlation			
1/4	.7320	.5936	−.2186	2.2338		.5937
1/2	.51425	.49894	−.35940	1.4324		.5041
3/4	.41435	.46285	−.44743	1.2469		.5188
1	.29967	.49437	−.45822	1.2950		.5014
			With Allowance for First-Order Serial Correlation			
1/4	.69709	.53934	−.18110		−.21466	.6005
1/2	.57847	.62016	−.47432		.37304	.5562
3/4	.59639	.68414	−.70429		.64122	.6503
1	.42062	.42621	−.56829		.46609	.5871

desirable consequence of inducing first- and higher-order serial autocorrelation in the residuals as the period of differencing increases. The results were recomputed allowing for first-order autocorrelation, but this is insufficiently sophisticated for the three-quarter and annual regressions and in any case makes no allowances for the autocorrelation in the original errors which, as we have seen, appear to be present for the *U.K.* model. Indeed, it can be seen from the table that the parameter estimates for the United Kingdom vary less systematically than those for the United States. Nevertheless, if due allowance is made for error, Table 1 does reveal several patterns. With one or two exceptions for the *U.K.* data the estimates do change in the directions indicated above. The coefficient on the lagged dependent variable increases with h, although rather more than it should in the United States and rather less in the United Kingdom; this is presumably due to the differing autocorrelation patterns in the original residuals. Most notable, however, is the fact that, for both countries, the proportional decline in the coefficient on ρ is much larger than that on π and that the latter can almost entirely be explained by changes in α according to equation (22). As indicated above, the parameter β can be expected to move towards unity as h increases, since over longer observation periods, a greater proportion of actual real income changes will have been observed; this accounts for the fact that the coefficient on ρ declines by more than is accounted for by α. However, since no such effect is observed for the inflation term, there appears to be little or no detectable adjustment of expected to actual inflation. This is then consistent with results using distributed lags and moving averages.

Although it is possible that expected inflation rates are in reality insensitive to the actual rate of inflation (and a number of explanations could be advanced for this), the conclusion remains implausible. The most likely explanation of our results is that the assumption of the constancy of k, the desired consumption income ratio, may be invalid, especially towards the end of the period. The argument would be that for moderate rates of inflation, the desired savings ratio is constant, but that it has risen over the last few years as inflation has climbed to historically high levels. Any one of a number of explanations can account for this, see Section III below. The disequilibrium inflation effects are not denied by this argument, but if, for example, we look at equation (17), and if the true value of k dips sharply towards the end of the period, the model can compensate for this by preventing π^* from rising, hence keeping $\sigma(1 - k) + \phi\pi^*$ close to its true value. A fall in k can thus be mistaken for insensitivity of inflation expectations. Note that if this argument is correct, it casts doubt only on the extent to which unanticipated inflation caused the *secular* rise in the savings ratio, and, in any case, this is a phenomenon which has any number of plausible explanations. It offers no alternative explanation for the well-attested instantaneous relationship between savings and inflation which occurs for both countries throughout the sample period.

III. Some Alternative Explanations

There are a number of possible mechanisms linking inflation and the savings ratio which differ from that examined above. In this section, some of these are discussed in the light of our results and their plausibility assessed as alternative explanations.

The first alternative rests on a possible reinterpretation of equations (23) and (24). For neither country is the coefficient on the rate of growth of real income very different from that on the inflation variable. If the two parameters were identical, the model would relate changes in the savings ratio solely to changes in *money* income. At first glance, this may appear to give a simple alternative explanation of the results in terms of saving acting as a "shock absorber," taking up unanticipated changes in money income. But such an explanation does not stand up to detailed examination. If real consumption responds sluggishly to

changes in income, changes in the savings ratio will be related to changes in *real* not money income, while an individual who systematically relates his savings ratio to changes in money income will find that rising inflation will reduce his standard of living even if real income is constant. Such behavior is not simple to explain and requires theoretical justification.

Second, there is a distributional argument which points out that changes in the rate of inflation are likely to be associated with changes in the share of disposable income accruing to various groups, for example, in favor of young workers at the expense of retired groups. Direct evidence relevant to this is hard to come by, but at least as far as recent British experience is concerned where income appears to have been redistributed away from professional middle-class earners towards unionized workers, it is implausible to suppose that such changes have produced an increase in the saving ratio. To the extent that this redistribution has been towards those with low marginal propensities to save, our results *understate* the true inflation effect on saving.

Third, at high rates of inflation, storeable commodities are substituted for money as the latter becomes more expensive to hold. Note that this effect requires that the rate of inflation be anticipated and is fully consistent with our model; once expectations catch up with or exceed the actual rate of price inflation, the illusion effect works in reverse with goods appearing to be cheap at any price. It is possible that actual recent inflation has caused little hoarding because expected rates of inflation have responded so slowly to observed rates. Even if this is understated, it is likely that the anticipated return from purchasing durable or storeable commodities has been insufficiently high to compensate for their carrying costs and illiquidity.

Fourth, there are the income and substitution effects between present and future consumption induced by changes in the real rate of interest. In the long run, there is perhaps a Fisherian process whereby real rates of return are independent of the rate of inflation, but this clearly does not operate in the short run. Even so, there is no presumption that a reduction in real interest rates will increase the saving ratio. The substitution effect works in the opposite direction, that is, in favor of current consumption, and although the income effect reduces the real wealth of net creditors, there are many debtors, for example, holders of home mortgages, for whom net worth will be increased. In some models of the consumption function, consumers are assumed to try to maintain a target ratio of wealth or of liquid assets to income and given this, increased inflation will induce extra saving. But this result is derived entirely by assumption and it is not clear why consumers should behave in this way in the face of changing relative prices between present and future consumption. At high rates of inflation, liquid assets become relatively costly and it may well be optimal to hold less, not more. The net effect is clearly in doubt although this is not to say that a dominant negative wealth effect is incapable of explaining the observed positive effect of inflation on saving. But this is unlikely. Wealth effects in consumption functions although often significant are rarely of great magnitude and once again, their operation requires time to allow some degree of anticipation by consumers. They can thus hardly explain the large and immediate effects which fluctuations in inflation appear to induce upon saving.

Fifth, and perhaps most convincing, are the uncertainty effects of inflation. These have been emphasized in recent empirical work on inflation effects in a series of papers by Thomas Juster and Paul Wachtel (1972a, b), and Juster and Lester Taylor. Since the results of these papers complement my own, it is worth dwelling on the arguments advanced. Juster and Wachtel argue that high rates of inflation have been associated historically with variable rates of inflation, so that if money income is not expected to match this variation, real income will be subject to greater uncertainty in times of high inflation. This

idea receives some support from survey evidence that inflation depresses consumer confidence. The higher uncertainty is then reflected in higher savings ratios as consumers seek to protect themselves against instability. For a theoretical treatment of these arguments see, for example, the paper by Frank Hahn.

In terms of our model, this would make k a function of the variance of real income, which, it is argued, is itself a function of the rate of inflation. Clearly, this possibility can be tested in future work. If we accept the validity of the argument, there is no difficulty in explaining the fact that current saving ratios are very high. The difficulties occur elsewhere. In Section II we noted a significant inflation effect in both countries over the period 1955–65 during which there was no major shift in the rate of inflation. Throughout the period, saving responded to price increases even when the latter were consistent with random fluctuations in a series of constant mean and variance. The mechanism linking inflation and saving thus seems capable of operating directly and does not necessarily require accompanying changes in uncertainty. However, Juster and Taylor find a measure of the variance of price expectations *across* consumers to be significant in their consumption function. This cross-section measure may not be a good proxy for uncertainty about price changes through time and it seems quite likely, as Juster and Taylor note, that this term reflects aggregation phenomena.

Otherwise, the results of the two Juster and Wachtel papers are entirely consistent with my model. They use the Michigan Survey Research Center series for price expectations so that they can directly separate out effects of anticipated and unanticipated inflation. They find that *unanticipated* inflation has a strong positive effect on saving with a negative effect on nondurable purchases and services; durable goods are relatively unaffected. *Anticipated* inflation increases nondurable purchases and services, and reduces saving and durable purchases. However, they note that inflation is rarely fully anticipated and it appears that the anticipated inflation series bears little or no relationship to actual inflation. In consequence, the apparent effects of anticipated inflation are open to possible reinterpretation, whereas all their other results are exactly as predicted by my model.

One set of results which at least appear to be inconsistent with mine and with those of Juster et al. have been published by William H. Branson and Alvin K. Klevorick. They estimate a "money illusion consumption function" of the form

$$\log c_t^* = \beta_0 + \sum_{i=0}^{I} \gamma_i \log y_{t-i}^* + \sum_{j=0}^{J} \delta_j \log w_{t-j}^* + \sum_{k=0}^{K} \eta_k \log P_{t-k} + \epsilon_t \tag{25}$$

where c_t^*, y_t^*, and w_t^* are real nondurable consumption, real income, and real wealth, respectively; P_t is the price level; and I, J, and K give the maximum length of distributed lag responses. This equation, apart from the inclusion of wealth, can quite reasonably be regarded as an approximation to equation (13), with the lags on y^* and P being convolutions of the expectational lags and those arising from the correction of previous mistakes. The surprising result is that Branson and Klevorick find $\eta_k > 0$ for all k, indicating a *negative* effect of inflation on saving. This contradicts my results, and it contradicts the three papers of Juster et al. One can only suppose that differences in data used are responsible for these results.

Finally we may refer to recent work done by the Bank of England. Their results cover the United Kingdom alone and relate to consumption functions not unlike those used by Branson and Klevorick. Again using Almon polynomials for lag structures, they find evidence consistent with the existence of an inflation effect on the savings ratio. After changes in real income have been allowed for, the immediate response of consumption to a rise in prices is negative, this becomes positive after one or two quarters, the sign of the ultimate response not being entirely clear. This result, which can

be replicated on *U.S.* data, is entirely consistent with my hypothesis, and would seem to tell against simple wealth effects. All these areas require a good deal more careful research.

APPENDIX

The Derivation of the Discrete Time-Estimating Equation

Write equation (20) in the form

$$\dot{z}_t = -\sigma z_t + \dot{x}_t + b + \zeta_t$$

where $z = (s/y)$, $\dot{x} = (1 - \beta)\rho - \phi\pi = \partial/\partial t\{(1 - \beta)log(y^*_t/P^*) - \phi log P\}$, $b = \{\sigma(1 - k) - (1 - \beta)\rho^0 + \phi\pi^0\}$, a constant, and ζ_t is a pure noise process to allow for errors in the model. A superimposed dot denotes differentiation with respect to time t. Integrating over the interval [t, t $-$ h],

$$z_t - z_{t-h} = (1 - e^{-\sigma h})(\sigma^{-1}b - z_{t-h}) - \int_0^h e^{-\sigma h}\frac{\partial x_{t-\theta}}{\partial\theta}d\theta + u_t$$

since $\dot{x} = -\partial x/\partial\theta$ and $u_t = \int_0^h e^{-\sigma\theta}\zeta_{t-\theta}d\theta$. Integrating by parts

$$z_t - z_{t-h} = (1 - e^{-\sigma h})(\sigma^{-1}b - z_{t-h}) + x_t - e^{-\sigma h}x_{t-h} - \sigma\int_0^h e^{-\sigma\theta}x_{t-\theta}d\theta + u_t$$

Phillips uses the approximation

$$x_{t-\theta} \doteqdot x_t - \theta\{x_{t-2h} - 4x_{t-h} + 3x_t\}/2h + \theta^2\{x_t - 2x_{t-h} + x_{t-2h}\}/2h^2$$

which allows explicit integration. Carrying this out and rearranging gives

$$z_t - z_{t-h} = (1 - e^{-\sigma h})(\sigma^{-1}b - z_{t-h}) + \alpha_1(x_t - x_{t-h}) + \alpha_2(x_{t-h} - x_{t-2h}) + u_t$$

where α_1 is given by α in equation (22) of the text and α_2 is

$$\alpha_2 = -\frac{1}{2\sigma h} + \frac{1}{(\sigma h)^2} - \left(\frac{1}{2\sigma h} + \frac{1}{(\sigma h)^2}\right)e^{-\sigma h}$$

An idea of the order of magnitude of α_2 can be gained by expanding the exponential; ignoring terms smaller than $(\sigma h)^2$, $\alpha_2 \doteqdot \sigma h/12 - (\sigma h)^2/24$. Since $\sigma h \doteqdot 1/4$, α_2 is small enough to be ignored. Hence, setting $\alpha_1 = \alpha$, $\alpha_2 = 0$, and substituting for z, b, and x, we reach equation (21) of the text.

REFERENCES

W. H. Branson, and A. K. Klevorick, "Money Illusion and the Aggregate Consumption Function," *Amer. Econ. Rev.*, Dec. 1969, *59*, 832–50.

Milton Friedman, *A Theory of the Consumption Function*, Princeton 1957.

F. H. Hahn, "Savings and Uncertainty," *Rev. Econ. Stud.*, Jan. 1970, *37*, 21–24.

F. T. Juster and P. Wachtel, (1972a) "Inflation and the Consumer," *Brookings Papers*, Washington 1972, *1*, 71–114.

——— **and** ———, (1972b) "A Note on Inflation and the Saving Rate," *Brookings Papers*, Washington 1972, *3*, 765–78.

——— **and L. D. Taylor,** "Towards a Theory of Saving Behavior," *Amer. Econ Rev Proc.*, May 1975, *65*, 203–09.

P. C. B. Phillips, "The Estimation of Some Continuous Time Models," *Econometrica*, Sept. 1974, *42*, 803–23.

Bank of England, "The Personal Saving Ratio," *Bank of England Quart Bull.*, Mar. 1976, *16*, 53–73.

U. K. Central Statistical Office, *Economic Trends*, London 1975.

U.S. Office of Business Economics, *Surv. Curr Bus.*, Washington 1975.

Price Behavior in U.S. Manufacturing: An Empirical Analysis of the Speed of Adjustment

By LEONARD SAHLING*

In recent years, the speed of adjustment of prices has come to be regarded as a critical issue for macroeconomic analysis. Indeed, it has been argued (by James Tobin and by Axel Leijonhufvud, ch. 2, especially pp. 67–81) that the assumption of relatively inflexible prices is the sine qua non of the Keynesian macroeconomic paradigm. To the extent that prices are slow to adjust to economic shocks, the burden of short-run adjustment will fall on quantity. Changes in the rate of production generate corresponding fluctuations in aggregate demand, and these fluctuations are amplified by multiplier-accelerator feedbacks. Hence, cumulative expansionary or contractionary episodes occur in the economy not just when prices are unvaryingly rigid, but whenever they are relatively sluggish in adjusting. The speed of adjustment of prices is thus an important empirical issue, and the purpose of this study is to develop an econometric procedure for measuring just how quickly they do adjust.

In this study, price behavior is cast in terms of a choice-theoretic model. Its major premises are that firms face an uncertain demand for their output and that their principal objective is to maximize the expected value of profits. Accordingly, their pricing decisions are directed toward effecting a balance between expected marginal cost and expected marginal revenue. Both incremental concepts, however, embody an implicit time dimension, with marginal cost and marginal revenue varying as the time horizon is changed. As it is not known a priori what time horizon firms actually use, a formal distinction is drawn between short- and long-run marginal costs, and both elements are provisionally included in the model of aggregate price formation. In this context the key concern is to determine the extent to which prices reflect temporary differences between short- and long-run marginal costs. This provides a basis for gauging the speed of adjustment of prices: the greater the responsiveness of prices to these temporary differences between short- and long-run marginal costs, the faster the speed of adjustment of prices.

This marginalist model is used to analyze aggregate price behavior in the *U.S.* nonfood manufacturing sector from 1955 to mid-1971. The empirical analysis is divided into two parts. First, a general form of the model is estimated under the working assumption that prices reflect both short- and long-run marginal costs. Second, the estimated model is then used to evaluate the degree to which prices are adjusted in response to differences between short- and long-run marginal costs. Within the general model, different degrees of short-run price adjustment—that is, full, partial, or none—impose alternative sign restrictions on certain coefficients. Each of the three specifications corresponds, then, to a distinct hypothesis about the speed of adjustment of prices. An econometric testing procedure is applied to determine which of them most accurately describes actual pricing behavior. According to the empirical results, the hypothesis of full adjustment of prices is rejected, but the other two hypotheses cannot be rejected. Although no firm conclusions can be drawn about whether prices are adjusted in part or not at all in response to the temporary differences between short-

*Economist, Federal Reserve Bank of New York. I would like to thank K. Hurley and M. Arak for their comments and suggestions on earlier drafts. I am indebted to R. Stein and P. Cahn for their excellent research assistance. Neither the Federal Reserve Bank of New York nor the Federal Reserve System necessarily concurs with the views in this paper.

and long-run marginal costs, the empirical evidence does seem to give the edge to the hypothesis of partial adjustment of prices. In any event, based on this evidence, the one clear-cut conclusion is that these output prices have tended to be less than fully flexible in the short run—in conformity with the Keynesian paradigm.

I. The Model

A. *Long-Run and Short-Run Pricing Decisions*

To begin, it is assumed that the three-factor Cobb-Douglas production function is an accurate characterization of the technological conditions in the manufacturing sector:

$$X_t = C_0 e^{\delta t}[L_t^{\alpha_1} V_t^{\alpha_2} K_t^{\alpha_3}] \tag{1}$$

where X_t is the output produced during period t; L_t is the labor input (man-hours) during period t; V_t is the materials input during period t; K_t is the stock of capital at the end of period t; and C_0 is a constant. Because of the lack of a precise measure of capital utilization, the flow of capital services has been taken to be proportional to the stock of capital. The materials input consists of those goods that originate outside of manufacturing and are purchased by manufacturers for further processing. This production function also assumes malleable capital, unitary elasticities of substitution among factors, and Hicks-neutral technical change equal to δ percent per quarter.

Firms face a stochastic demand. Quantity demanded per period (Q_t^d), from their perspective, can be regarded as the product of two multiplicative but unobservable components. Accordingly, all factors affecting demand are grouped into two broad categories—deterministic and stochastic:

$$Q_t^d = [Z_t P_t^{\theta}][e^{v_t}] \tag{2}$$

The first bracketed term covers the systematic factors which, in theory, predictably affect the demand for output. One of the main ones, for this study, is the responsiveness of demand to the prices charged by firms; it is given by P_t^{θ}, where P is the price and θ is a parameter measuring the elasticity of demand to price, that is, a parameter which neither varies over time nor over the business cycle. Only through this mechanism are firms able to exercise partial control over quantity demanded. The other systematic factors influencing demand are beyond their individual control—for example, general business conditions, tax rates, and the prices of other goods. These other factors are represented collectively by the shift parameter (Z_t) and will be referred to as exogenous demand. The exponential term in the second pair of brackets encompasses the nonsystematic stochastic disturbances that impinge on demand. It is assumed that the exponent v is distributed normally with mean zero and variance σ^2, and that expectations about exogenous demand (Z_t^+) are formed independently of the stochastic component. Hence, letting the expected-value operator be denoted by E, the expected value of demand is equal to

$$E(Q_t^d) = Z_t^+ P_t^{\theta} e^{\sigma^2/2} \tag{3}$$

Central to the model of how firms go about determining their prices are three key behavioral postulates: 1) Firms behave as monopolists and retain decision-making control over their own prices.[1] 2) The objective of firms is to maximize the expected value of their prospective profits. 3) The time horizon over which firms strive to attain their objective extends far enough into the future that all factor inputs can be regarded as being variable. Within this setting, and under the simplifying assumption that the quantities produced and demanded must in the long run be equal, the objective of firms can be posed in terms of constrained maximization:

[1]Not all manufacturing industries qualify as monopolistic under the usual textbook definition. However, Kenneth Arrow has argued that the combination of uncertainty, imperfect knowledge, and disequilibrium imparts some degree of inelasticity to the demand for the output of *all* firms, even those in otherwise objectively competitive markets.

$$(4) \quad \underset{P,L,V,K}{\text{Max}}\, \Pi = P_t E(Q_t^d) - W_t^+ L_t - M_t^+ V_t - R_t^+ K_t - \lambda_{1t}[E(Q_t^d) - C_0 e^{\delta t} L_t^{\alpha_1} V_t^{\alpha_2} K_t^{\alpha_3}]$$

where W_t is the wage rate per man-hour in period t, M_t is an index of materials prices in period t, and R_t is the rental price of capital in period t. In this equation, the prospective values expected over the long-term horizon have been denoted by a superscript +; these expectations are assumed to be independent of each other, as well as of the stochastic component of demand. When this equation is differentiated with respect to price, factor inputs, and the Lagrangian multiplier, the first-order conditions for maximizing expected profits are obtained.

Among these necessary conditions is the familiar one requiring equality between expected marginal revenue and marginal cost. With a random disturbance term appended to it to allow for unaccountable factors such as "managerial inertia, ignorance, etc." (Dorothy Hodges, p. 722), this first-order condition can be written as

$$(5) \quad \left(1 + \frac{1}{\theta}\right) P_t^* = \lambda_{1t} e^{u_t'}$$

Herein, λ_{1t} refers to expected long-run marginal cost, and the disturbances are assumed to be distributed independently of the stochastic component of demand. The other first-order conditions relate to the optimal factor inputs, denoted as L_t^*, V_t^*, and K_t^*, and the constraint that production be equal to expected quantity demanded:

$$(6a) \quad L_t^* = \lambda_{1t}\alpha_1 E(Q_t^d)/W_t^+$$

$$(6b) \quad V_t^* = \lambda_{1t}\alpha_2 E(Q_t^d)/M_t^+$$

$$(6c) \quad K_t^* = \lambda_{1t}\alpha_3 E(Q_t^d)/R_t^+$$

$$(6d) \quad E(Q_t^d) = C_0 e^{\delta t} L_t^{*\alpha_1} V_t^{*\alpha_2} K_t^{*\alpha_3}$$

Obtaining the decision rule for the optimal long-run price is now a fairly simple matter. An explicit expression for λ_{1t} can be derived from equations (3) and (6a–d) and then substituted into (5); next, the terms in the resulting expression can be rearranged so that P stands alone on the left-hand side:

$$(7) \quad P_t^* = C_1 e^{-(\delta/\gamma)t} \cdot [(W_t^+)^{\alpha_1}(M_t^+)^{\alpha_2}(R_t^+)^{\alpha_3}]^{(1/\gamma)}(Z_t^+)^{(1-\rho)/\gamma} e^{u_{1t}}$$

wherein C_1 is a constant and the following definitions hold:

$$\rho \equiv (\alpha_1 + \alpha_2 + \alpha_3)$$
$$\gamma \equiv (\rho - \theta(1 - \rho))$$
$$u_{1t} \equiv (\rho/\gamma)u_t'$$

In what follows, it will be assumed that the elasticity of the *aggregate* demand for manufactured nonfood goods to prices is negative, i.e., $\theta < 0$.[2] Thus, the stability of the model requires that $\gamma > 0$, for then and only then is the elasticity of total cost with respect to output $(1/\rho)$ greater than the corresponding elasticity of revenue $(1 + 1/\theta)$. Were this condition not met, it would be profitable for firms to expand (contract) output as much as possible when marginal revenue is above (below) marginal cost.

A different set of decision rules will apply in the short run, here assumed to be no longer than a quarter year. The differences stem from two intrinsic features of the short run: First, no matter how well firms can forecast, their predictions of prospective demand will still be inexact, owing to the random vagaries of demand. In response to such forecasting errors, firms will have to implement intraperiod accommodative adjustments in either price or quantity—or both for that matter. Second, neither labor nor capital is completely flexible in the short run. To the extent, then, that output is varied in the short run, the inflexibility of the "fixed" factors opens up a breach between short-run and expected long-run marginal costs.

In modeling the short-run behavior of firms, it is assumed that investment goods acquired during a given period do not begin to yield capital services until the next period, and that the actual input of labor is adjusted toward the desired value, denoted as $\hat{L}$, at a constant geometric rate:

[2]For empirical evidence that supports this assumption, see the articles by William Branson and Alvin Klevorick, and by James Pierce and Jared Enzler.

$$(8) \qquad (L_t/L_{t-1}) = (\hat{L}_t/L_{t-1})^{\eta}$$

where $0 \leqq \eta \leqq 1$.

For profit-minded firms, the optimization problem that is faced in the short run can be expressed as:

$$(9) \quad \underset{P,L,V}{\text{Max}}\, \Lambda = P_t Q_t^d - L_t W_t - M_t V_t - R_t K_{t-1} - \lambda_{2t}[Q_t^d - C_0 e^{\delta t}(L_t^{\alpha_1} V_t^{\alpha_2} K_{t-1}^{\alpha_3})]$$

In the short run, as depicted here, the size of the capital stock is not a discretionary variable, and equality between the quantities produced and demanded is achieved through variations in both price and quantity produced.[3] After differentiating equation (9), the first-order conditions can be solved for the short-run program for maximizing profits:

$$(10a) \qquad \hat{P}_t = (1 + 1/\theta)^{-1}\lambda_{2t}$$

$$(10b) \qquad \hat{L}_t = \alpha_1 \lambda_{2t} Q_t^d / W_t$$

$$(10c) \qquad \hat{V}_t = \alpha_2 \lambda_{2t} Q_t^d / M_t$$

$$(10d) \qquad Q_t^d = C_0 e^{\delta t}[\hat{L}_t^{\alpha_1} \hat{V}_t^{\alpha_2} K_{t-1}^{\alpha_3}]$$

Actually, the short-run price $\hat{P}$ will not be optimal unless the labor input is totally flexible. But, by combining equations (2), (8), and (10a-c) with the short-run production function, the *feasible* optimal short-run price, denoted as P' (and distinct from P unless $\eta = 1$), can be obtained:[4]

$$(11) \quad P_t' = C_2 e^{-(\delta/\tau)t}\{[W_t^{(\eta\alpha_1/\tau)} M_t^{(\alpha_2/\tau)}] [L_{t-1}^{-(1-\eta)\alpha_1} K_{t-1}^{-\alpha_3}]^{1/\tau} Z_t^{(1-\kappa)/\tau}\} e^{\nu_t'}$$

where C_2 is a constant term and the following definitions hold:

$$\kappa \equiv (\eta\alpha_1 + \alpha_2)$$
$$\tau \equiv (\kappa - \theta(1 - \kappa))$$
$$\nu_t' \equiv ((1 - \kappa)/\tau)\nu_t$$

Note that $\tau > 0$ so long as $\gamma > 0$ (see equation (7)), and that κ measures the returns to scale that govern short-run operations.

Suppose, for the time being, that firms do revise their prices in the short run to be equal to P'. (Some reasons why they might not behave in this manner will be discussed in the next subsection.) To facilitate the analysis, let the question of how expectations are formed be postponed for now, and assume that the forecasts of wages, materials prices, and exogenous demand are on average equal to the actual values, i.e., $W = W^+$, $M = M^+$, and $Z = Z^+$. Given this assumption, and upon combining the derivations of equations (6a–d), (7), and (11), the terms can be rearranged in order to show how short-run prices stand in relation to optimal long-run prices:

$$(12) \quad (P_t'/P_t^*) = (L_t^*/L_{t-1})^{((1-\eta)\alpha_1)/\tau}(K_t^*/K_{t-1})^{(\alpha_3/\tau)} e^{u_{3t}} = (\lambda_{2t}/\lambda_{1t}) e^{-u_t'}$$

According to this relationship, whether the optimal short-run price is above or below its long-run counterpart depends upon the deviations between the optimal and actual inputs of the fixed factors of production. This is a direct implication of the model, which depicts firms as adjusting output more quickly than they do their inputs of labor and capital. Within the context of the model, such speedy adjustments in output can only be achieved through extraordinarily large increases or decreases, as the case may be, in raw materials and, to a lesser degree, in man-hours. By the same reasoning, moreover, $P' \rightarrow P^*$ over time, as firms both revise their assessments of prospective demand conditions and adjust their fixed inputs accordingly.

[3]Alternatively, in accommodating demand, firms could vary unfilled orders or finished-goods inventories, while they maintain price and output at constand levels. Indeed, the distinction between firms that behave in this manner and those that do not plays a fundamental but tacit role in the pricing model that is formulated in the next subsection.

[4]The key step in this derivation consists of setting up the short-run production function as follows:

$$(10d') \quad Q_t^d = C_0 e^{\delta t}[L_{t-1}(\hat{L}_t/L_{t-1})^{\eta}]^{\alpha_1} V_t^{\alpha_2} K_{t-1}^{\alpha_3}$$

which holds since $(L/L_{-1}) = (\hat{L}/L_{-1})^{\eta}$ by equation (8). After equations (2) and (10a-c) have been substituted into (10d'), the resulting expression can be simplified so that P' stands alone on the left-hand side.

B. *Industrial Price Formation: A Marginalist Model*

Actual prices would be equal to P' if firms were unencumbered profit maximizers. In practice, however, there could be certain frictions which impede firms from fully adjusting their prices. For example, Armen Alchian has pointed out that firms might well choose to hold their prices relatively constant in the face of short-run fluctuations in demand as a service to their customers. Presumably, to the extent that buyers thereby incur lower average search costs, sellers will exact a quid pro quo by charging a relatively higher average price. But producers face another impediment. Administrative and information gathering costs will be incurred when firms try to decide whether and by how much to vary their prices. Robert Barro has contended that these costs "... can reasonably be described as a lump-sum amount, independent of the size or direction of [price] adjustment" (p. 21). He then goes on to argue that a firm will only adjust its price when the foregone profits from not charging the optimal price exceed the threshold established by the lump sum costs.

Now as an extension of Barro's argument, assume that these lump sum amounts vary in size among firms. Let firms be arrayed along a scale according to the size of their lump sum amounts. In general, firms at the high end of the scale will infrequently adjust their prices, whereas those at the low end will tend to "fine-tune" their prices. Given this continuum, the following expression encompasses the full range of feasible outcomes for aggregate pricing behavior:

$$P_t = (P_t')^{\epsilon}(P_t^*)^{(1-\epsilon)} \tag{13}$$

where P_t refers to the actual price and $0 \leqq \epsilon \leqq 1$. In effect, the parameter ϵ can be regarded as an *ex post* measure of the elasticity of actual prices to the discrepancies between short- and long-run marginal costs.

Equation (13) can be used to formulate an eclectic model which embodies the two optimal pricing rules as special cases. Upon substituting equations (7) and (11) into (13), and then combining like terms under the assumption that the forecasts of wages, materials prices, and exogenous demand are on average equal to the actual values, the resulting expression can be simplified to the basic model of aggregate price formation:

$$P_t = A_0 e^{A_1 t} \cdot [(W_t^+)^{A_2}(M_t^+)^{A_3}(R_t^+)^{A_4}(Z_t^+)^{A_5}] \cdot [(L_{t-1})^{A_6}(K_{t-1})^{A_7}] e^{u_{4t}} \tag{14}$$

where the coefficients are equal to:

$$A_1 = -(\delta/\gamma)[(1 - \epsilon) + \epsilon(\gamma/\tau)] < 0$$
$$A_2 = (\alpha_1/\gamma)[(1 - \epsilon) + \epsilon(\eta\gamma/\tau) > 0$$
$$A_3 = (\alpha_2/\gamma)[(1 - \epsilon) + \epsilon(\gamma/\tau)] > 0$$
$$A_4 = (\alpha_3/\gamma)(1 - \epsilon) \geqq 0$$
$$A_5 = [(1 - \rho)/\gamma][1 + (\epsilon/\tau)[(\rho - \kappa)/(1 - \rho)]] \ ?\ 0$$
$$A_6 = -\epsilon[(1 - \eta)\alpha_1/\tau] \leqq 0$$
$$A_7 = -\epsilon(\alpha_3/\tau) \leqq 0$$

The sign restrictions are intuitively sensible, and are all based on earlier assumptions about the behavioral parameters: α_1, α_2, $\alpha_3 > 0$; $\theta < 0$; δ, γ, $\tau > 0$; and $0 \leqq \epsilon$, $\eta \leqq 1$. The ambiguity about the sign on the exogenous-demand term arises because of the absence of prior information about returns to scale in the long run.[5] Yet, as a practical matter, it would indeed be surprising if there were increasing returns to scale at work in the aggregate and over the long run. Hence, given the a priori expectation of constant or decreasing economies of scale, shifts in the aggregate demand curve will either have no effect at all or tend to "pull" prices in the same direction, i.e., $A_5 \geqq 0$.

In an empirical application, the arguments in equation (14) are sure to involve distributed lag (*DL*) effects of different durations. In some cases, these *DL* effects will reflect expectation-formation processes. It is the *expected* values of wages, materials

[5]Given the definition of A_5, it can be shown that $A_5 > 0$ if $\rho < 1$; and that if $\rho > 1$, then $A_5 \lesseqgtr 0$, as $(\epsilon/\tau) \gtrless [(\rho - 1)/(\rho - \kappa)]$.

prices, and the rental price of capital that appear in the model, not the actual values. Assuming that these expectations are formed on the basis of past and present experience, the expected factor-price arguments may then be replaced by lag distributions of current and past actual values:

$$W_t^+ = \Pi(W_{t-i})^{\beta_{2i}}$$
$$M_t^+ = \Pi(M_{t-j})^{\beta_{3j}}$$
$$R_t^+ = \Pi(R_{t-g})^{\beta_{4g}}$$

Accordingly, the sum of the *estimated DL* coefficients will be equal to the product of two elements: the sum of the expectations coefficients; and the corresponding coefficient (A_i) from equation (14). Following Franco Modigliani and Richard Sutch, these expectations are taken to be blends of regressive and extrapolative processes. Hence the *DL* coefficients need not take on a simple geometric form, nor will they necessarily sum to unity. This in turn means that the price equation will be underidentified. Yet, as long as the sums of the expectation coefficients are positive, the sums of the estimated *DL* coefficients must still conform to the inequality restrictions given in connection with equation (14).

Expected exogenous demand (Z^+) cannot be modeled in exactly the same way, since there is not an empirically observable measure upon which to anchor it. However, the schematic factorization of demand given in equation (2) does suggest that the exogenous component will be closely related to total quantity demanded. Taking advantage of this theoretical association, a combination of current and past values of quantity demanded can be used as a stand-in for expected exogenous demand:

$$Z_t^+ = [\Pi(Q^d_{t-s})^{\beta_{5S}}]$$

Once again, the pattern of these coefficients will reflect—though less precisely than before—the underlying extrapolative and regressive processes, and the coefficients will presumably sum to a positive value, i.e., $\Sigma\beta_{5S} > 0$. Alternatively, taken at face value, this autoregressive-like expression implies that, in forming these expectations, firms disregard the effect of past changes in prices on prospective demand. Given the formidable problems that firms face in forecasting, this characterization may not be too much of an oversimplification.

There is another reason for expecting to observe distributed lag responses. Within the manufacturing sector, firms are highly interrelated in the sense that the outputs of some firms are inputs of others. Given this intricate network of "input-output" linkages, the decision by even a few firms to raise their prices will set off a chain reaction, for other firms will then react by raising their prices. And so the process will continue in multiplier-like fashion. Note that this process can be triggered by any one of the independent variables in the model, including those which enter the model only as a result of the hypothesized influence of the gap between short- and long-run costs on prices (i.e., labor and capital). Moreover, once triggered, the adjustment mechanism might well take longer than a quarter year before it unfolds fully.

For these reasons, then, each of the arguments in equation (14) was replaced by a lag distribution of actual values. After taking logarithms, the resulting expression was transformed into first differences, in order to filter the time trends from the data series. Thus, the following form of the pricing model is the one that will be estimated:

$$(15) \quad \Delta \ln P_t = b_1 + \underset{(+)}{\sum_0^I b_{2i}\Delta \ln W_{-i}} + \underset{(+)}{\sum_0^J b_{3j}\Delta \ln M_{-j}} + \underset{(+\text{ or }0)}{\sum_0^G b_{4g}\Delta \ln R_{-g}} + \underset{(+)}{\sum_0^S b_{5s}\Delta \ln Q^d_{-s}} + \underset{(-\text{ or }0)}{\sum_1^E b_{6e}\Delta \ln L_{-e}} + \underset{(-\text{ or }0)}{\sum_1^F b_{7f}\Delta \ln K_{-f}} + u_{5t}$$

Listed below each of the lag distributions is the corresponding sign requirement. These requirements are, of course, consistent with those that apply to equation

(14). In three cases, the inequality restrictions are supposed to hold independently of the extent to which prices are varied in response to the short-run differences in their marginal costs: i.e., Σb_{2i}, $\Sigma b_{3j} > 0$, and $\Sigma b_{5s} \geqq 0$, for $0 \leqq \epsilon \leqq 1$. In the other three cases, however, the restrictions do depend critically on the value of ϵ:

$$(16)\quad \epsilon = 0: \Sigma b_{4g} > 0 \text{ and } \Sigma b_{6e} = \Sigma b_{7f} = 0$$
$$0 < \epsilon < 1: \Sigma b_{4g} > 0 \text{ and } \Sigma b_{6e}, \Sigma b_{7f} < 0$$
$$\epsilon = 1: \Sigma b_{4g} = 0 \text{ and } \Sigma b_{6e}, \Sigma b_{7f} < 0$$

Determining which of these outcomes, if any, best describes aggregate price formation is an empirical task—one that is undertaken in Section III.

II. Econometric Analysis

Quarterly observations, unadjusted for seasonal variation, were used to estimate the model.[6] The sample period ranges from 1953-I to 1971-II. Later data were not used because the Phase I price controls were instituted in August 1971; data limitations prevented the sample period from being extended in the other direction as well.

The data used cover the nonfood manufacturing sector of the United States, that is, the manufacturing sector excluding the food processing firms that comprise the food and beverage industry.[7,8] Wholesale price data were used to measure the prices of both nonfood crude materials as well as manufactured nonfood products. Wages were defined as the average hourly earnings of production workers, excluding the effects of overtime pay and of interindustry shifts in the composition of employment. The input of labor was defined as the total man-hours of production workers, and the capital stock includes both plant and equipment. The other variables, quantity demanded and the rental price of capital, are discussed below in the text.

Several stumbling blocks had to be overcome before the model could be estimated. One such impediment arises in connection with the impact of demand on prices. Within equation (15), it is maintained that firms vary their prices in response to changes in quantity demanded Q^d, defined here as the rate of constant-dollar new orders.[9] Yet causation almost surely runs in the opposite direction as well. Hence, to avoid simultaneity bias it was decided to estimate the model using an instrumental-variables technique. Contemporaneous values of real new orders were accordingly replaced with the predicted values obtained

[6]Quarterly dummy variables had initially been included in the regression equation to allow for seasonal variations in the data. These variables were subsequently dropped from the analysis, however, because they had such a minimal and statistically insignificant impact.

[7]A data appendix describes how the variables were constructed, gives a bibliography of the data sources, and lists the actual quarterly time series. Copies of this appendix are available from the author upon written request.

[8]There were two a priori reasons for excluding the food and beverage industry (i.e., industry #20 in the Standard Industrial Classification Code). First, it seemed likely that food processing firms behaved differently in the short run than do the other manufacturing firms. One of the key premises of the pricing model is that firms react to short-run variations in quantity demanded by adjusting both price and output. Yet, in the short run, the output of food processing firms is largely governed by the highly unstable supply conditions in the agricultural sector. Hence, these firms would perforce be much more reliant on short-run variations in prices than are the other firms.

The second reason was to avert an aggregation problem which would otherwise have arisen in connection with the index of the wholesale prices of raw materials used by manufacturers. In constructing this index, the Bureau of Labor Statistics (*BLS*) uses fixed weights equal to the relative gross value of shipments in the base year. In fact, the proportion of raw foodstuffs to the total raw materials input to manufacturing is much bigger than the ratio of manufactured foods and beverages to the total output of manufacturing. At the same time, the relatively large fluctuations in raw food prices are independent of the movements in the prices of other raw materials. Thus, if the prices of raw food stuffs had not been deleted, the index of raw materials prices would then have been both generally unrepresentative and disproportionately noisy.

[9]This hypothesis is a distinguishing feature of the model developed in this study. In contrast, almost all other empirical models have postulated that changes in prices are related either to the level of excess demand or to cumulated past rates of excess demand. See, for example, George de Menil's study and William Nordhaus's review of other studies (Table 3, pp. 36–37).

from a regression of real new orders on a set of instrumental variables.[10]

Nor can the assumption of unidirectional causation be validly applied to the rental price of capital R. Following Roger Waud, p. 408, this variable was defined as:

$$R_t = P_t^k(r_t + d) \tag{17}$$

Here P^k is an index of the prices of plant and equipment purchased by manufacturers; r_t is the cost of capital, defined as the (quarterly) interest rate on long-term *U.S.* Treasury bonds; and d is the quarterly rate of depreciation of the stock of fixed capital goods owned and operated by manufacturers. Under this definition, the supply prices of capital goods (P^k) are one of the primary components of the rental cost of capital; but many of these prices are also components of the wholesale prices of manufactured goods, that is, the dependent variable for this study. Hence, the contemporaneous value of the rental price of capital is correlated, by definition, with the residual of the price equation. While the one sure way out of this conundrum would be to disaggregate further, the alternative solution adopted here was to assume that the distributed lag impact of R_t on prices begins with a one-period delay—a workable solution only if the residuals are serially independent.

The third and last major impediment to estimating the model was the lack of a priori information concerning the lengths of the distributed lags. Even if it were assumed that the distributed lags were no longer than two or three years, estimating all possible forms of equation (15) would still have been an impossibly huge undertaking. I invoked instead an adaptation of the *ad hoc* search procedure devised by Branson and Klevorick. This is a two-step procedure that involves first setting all the lag lengths at four periods, and then experimenting systematically with other settings in order to find the one that minimizes the standard error of estimate. Almon's polynomial technique was applied in those instances where the length of the distributed lag exceeded four periods. (That is, in applying this procedure, the degree of the polynomial was fixed at three, and endpoint restrictions were *not* imposed.) The distributed lags were otherwise estimated freely. Also, because of data limitations and the manner in which the distributed lags were estimated, the start of the sample period had to be moved ahead from 1953-I to 1955-I.[11]

Selected estimation results are reported in Tables 1 and 2. In the initial regression, the lengths of the distributed lags had all been set at four periods, $I = J = S = 3$ and $E = F = G = 4$. This appears as regression 1-1 in Table 1, and it provides a standard of comparison for the final results. The corresponding individual distributed lag coefficients are listed in the top portion of Table 2. Next, guided in part by the initial regression results, the search was then extended over a grid of alternative values of I, J, etc. In all, over fifty regressions were estimated. The one with the lowest standard error of estimate was chosen as embodying the best estimate of the lag structure of the pricing model. This appears as regression 1-2 in Table 1; the corresponding distributed lag coefficients are listed in the bottom portion of Table 2. Note how well these estimated coefficients conform to the a priori sign requirements. Nothing in the search procedure guaranteed this outcome.

As a check on the model, the Chow te[...] was applied to see whether there had been [...]

[10] This instrumental-variables regression is described in detail in the data appendix. The key instrumental variables chosen by trial and error were the private nonfarm inventory-to-sales ratio, wholesale prices of nonfood raw materials used by manufacturers, the ratio of the market value of nonfinancial corporations to the reproduction cost of their physical assets, and the market yield on three-month Treasury Bills. At 0.72, the R^2 for this regression is low for the first stage of a two-stage estimation procedure, but it is high relative to comparable "reduced-form" models which have been used elsewhere to explain changes in *real GNP*.

[11] Data on new orders received by nonfood ma[...] facturers are available only from 1953-I. Wherea[...] had been decided at the outset to allow for a [...] tributed lag on real new orders of as long as two ye[...] the beginning of the sample period had to be mo[...] ahead two years to 1955-I.

TABLE 1—INSTRUMENTAL-VARIABLE ESTIMATES OF THE PRICING MODEL, TEXT EQUATION (15)

Independent Variable	Number of Lags	Regression Number 1-1	1-2	1-3
$\Delta \ln W$	$\sum^{I} b_{2i}$	3 0.510 (2.95)	2 0.365 (2.62)	2 0.405 (2.58)
$\Delta \ln M$	$\sum^{J} b_{3j}$	3 0.086 (1.32)	0 0.090 (4.76)	0 0.088 (4.39)
$\Delta \ln R$	$\sum^{G} b_{4g}$	4 0.357 (3.43)	8 0.485 (5.25)	
$\Delta \ln P^{k}$	$\sum^{G} b'_{4g}$			8 0.517 (4.65)
$\Delta \ln (r + d)$	$\sum^{G} b''_{4g}$			8 0.389 (2.34)
$\Delta \ln Q^{d}$	$\sum^{S} b_{5s}$	3 0.114 (3.26)	4 0.211 (7.37)	4 0.207 (6.31)
$\Delta \ln L$	$\sum^{E} b_{6e}$	4 −0.032 (0.68)	3 −0.133 (3.73)	3 −0.121 (2.86)
$\Delta \ln K$	$\sum^{F} b_{7f}$	4 0.018 (0.31)	3 0.047 (1.15)	3 0.031 (0.66)
Constant	b_1	−0.0056 (3.98)	−0.0064 (6.59)	−0.0066 (6.11)
$\bar{R}^2$		0.757	0.848	0.837
$SEE \times 10^2$		0.2485	0.1970	0.2034
DW		1.59	1.98	2.05

Note: Sample period of dependent variable 1955-I to 1971-II, t-statistics shown in parentheses.

structural shift in the coefficients over the sample period. Instability, in this sense, is often a symptom of a misspecified model. The sample period was divided into two parts. One part covered the period up to 1966-IV, and the other included all subsequent observations. (This dividing point was chosen because it marks the beginning, more or less, of the prolonged acceleration in inflation that culminated in the wage and price controls that were imposed in August 1971.) When regression 1-2 was reestimated over the two subperiods, the calculated value of Chow's F-statistic was found to be 0.92. Hence, the null hypothesis of equality between the sets of estimated coefficients for the two subperiods could not be rejected.

The model was subjected to another test. The question arises whether the significant coefficient of the rental price of capital in regression 1-2 is the result of spurious correlation. To see this, let definition (17) be generalized as:

$$\ln R_t = \phi_1 \ln P_t^k + \phi_2 \ln(r_t + d) \tag{18}$$

where ϕ_1, ϕ_2 are constants. Up to this point, equation (15) has been estimated under the tacit assumption that $\phi_1 = \phi_2$. This is the source of the problem—if indeed there is one. Whereas all output prices in the economy tend to move together, it is unclear whether a significant coefficient on the rental price of capital is reflecting anything beyond this common movement. In the extreme, the interest rate component of the rental price argument might not be contributing at all to the explanatory power of the regression model.

The aim is to test whether $\phi_1 = \phi_2$. Let the lag distribution on the rental price of capital in equation (15) be divided into

TABLE 2—DISTRIBUTED LAG PROFILES FOR THE COEFFICIENTS IN REGRESSIONS 1-1 AND 1-2

Regression 1-1: Initial Estimate

i	$\Delta \ln W_i$	$\Delta \ln M_i$	$\Delta \ln R_i$	$\Delta \ln Q_i^d$	$\Delta \ln K_i$	$\Delta \ln L_i$
0	0.471	0.096	-	0.017	-	-
	(3.77)	(3.54)		(0.85)		
-1	0.278	0.019	0.069	0.048	0.234	-0.011
	(2.25)	(0.63)	(1.54)	(3.43)	(1.90)	(0.41)
-2	-0.183	-0.021	0.049	0.036	-0.020	-0.026
	(1.33)	(0.75)	(0.99)	(2.61)	(0.15)	(0.99)
-3	-0.057	-0.008	0.131	0.013	-0.237	-0.016
	(0.41)	(0.29)	(2.75)	(0.96)	(1.99)	(0.68)
-4	-	-	0.108	-	0.041	0.021
			(2.29)		(0.42)	(1.02)
-5	-	-	-	-	-	-
-6	-	-	-	-	-	-
-7	-	-	-	-	-	-
-8	-	-	-	-	-	-
	0.510	0.086	0.357	0.114	0.018	-0.032
	(2.95)	(1.32)	(3.43)	(3.26)	(0.31)	(0.68)

Regression 1-2: Final Estimate

i	$\Delta \ln W_i$	$\Delta \ln R_i$	$\Delta \ln Q_i^d$	$\Delta \ln K_i$	$\Delta \ln L_i$
0	0.344	-	0.020	-	-
	(3.66)		(2.19)		
-1	0.215	0.057	0.063	0.238	-0.037
	(2.28)	(2.13)	(6.65)	(2.97)	(1.81)
-2	-0.193	0.091	0.064	0.022	-0.056
	(1.81)	(4.35)	(7.32)	(0.32)	(3.30)
-3	-	0.113	0.043	-0.213	-0.040
		(4.75)	(5.23)	(2.59)	(2.52)
-4	-	0.119	0.020	-	-
		(5.63)	(1.93)		
-5	-	0.107	-	-	-
		(5.31)			
-6	-	0.072	-	-	-
		(3.25)			
-7	-	0.010	-	-	-
		(0.47)			
-8	-	-0.083	-	-	-
		(3.25)			
	0.365	0.485	0.211	0.047	-0.133
	(2.62)	(5.25)	(7.37)	(1.15)	(3.73)

Note: t-statistics are shown in parentheses.

two parts:

$$\Sigma b_{4g} \Delta \ln R_{t-g} = \Sigma b'_{4g} \Delta \ln P^k_{t-g} + \Sigma b''_{4g} \Delta \ln (r_{t-g} + d) \tag{19}$$

Thus $\phi_1 = \phi_2$ if and only if $b'_{4g} = b''_{4g}$ for $g = -1, \ldots, -8$. To test the latter conditions, regression 1-2 in Table 1 was reestimated with the two components of the rental price of capital entered as separate arguments. The result is reported in column 3 of Table 1. The sums of squared residuals are equal to 1.82355×10^{-4} for regression 1-2, and to 1.77980×10^{-4} for regression 1-3. Use of the Almon Technique reduces the number of equality restrictions from eight to four. Under the null hypothesis that $\phi_1 = \phi_2$, the calculated value of F amounts to:

$$F(4, 43) = \frac{(1.82355 - 1.77980)/4}{1.77980/43} = 0.26$$

Clearly, the null hypothesis cannot be rejected. This means that the two components of the rental price of capital are contributing about equally to the overall explanatory power of the regression.

For all its strengths, regression 1-2 does feature one notable flaw. The three coefficients on the lagged capital-stock variables add up to a positive value, not a negative one as the model requires. Even though two of the three individual coefficients are significantly different from zero, they take on opposing signs and the sum turns out not to be significantly different from zero. This perplexing pattern could reflect some basic weakness either in the model or in the working hypothesis that prices are adjusted in response to differences between short- and long-run marginal costs. Both possibilities are tested in the next section. Alternatively, however, the problem could stem from how the capital stock has been measured. The capital-stock estimates used in this study (i.e., quarterly interpolations of annual estimates compiled by the Bureau of Economic Analysis) embody a number of assumptions. Among them is the assumption that two classes of fixed-capital goods made up of plant and equipment are perfect substitutes in the production process. This does not square with what empirical evidence there is on the matter. Perhaps, then, the widely scattered coefficients on the lagged capital-stock variables are signaling the inappropriateness of this tacit assumption. This conjecture could not be verified, however, because of the lack of quarterly data for the manufacturing sector which distinguish between plant and equipment investment.

III. An Empirical Analysis of the Speed of Adjustment of Prices

Now that the model has been estimated, it will be used to evaluate the speed of adjustment of prices. This will be done by determining to what extent prices are adjusted in response to the temporary differences between short- and long-run marginal costs: The greater the responsiveness of prices to these temporary differences, the faster the speed of adjustment of prices. Different degrees of short-run price adjustment—that is, full, partial, or none—impose alternative sign or zero restrictions on certain coefficients of the model. By the above criterion, each specification corresponds to a distinct hypothesis about the speed of adjustment of prices. An econometric procedure will be applied to determine which of these three hypotheses most accurately describes the actual pricing behavior of nonfood manufacturing firms.

Each of the three hypotheses about the speed of adjustment of prices corresponds to a particular specification of equation (15). Let H_A denote the hypothesis of full short-run flexibility. This is the case where firms fully adjust their prices in response to temporary differences between short- and long-run marginal costs, where $\epsilon = 1$. In terms of equation (15), this hypothesis would be accepted if and only if:[12]

$$H_A: \Sigma b_{4g} = 0 \text{ and } \Sigma b_{6e}, \Sigma b_{7f} < 0$$

Focussing only on the feasible portion of the sample space, as defined by the model,

[12]For details, see equation (16) and the accompanying text.

the only other tenable hypotheses correspond to the following two outcomes:

$$H_B: \Sigma b_{4g} > 0 \text{ and } \Sigma b_{6e}, \Sigma b_{7f} < 0$$
$$H_C: \Sigma b_{4g} > 0 \text{ and } \Sigma b_{6e} = \Sigma b_{7f} = 0$$

In the case (H_B:$0 < \epsilon < 1$), some firms adjust their prices in the short run, but others do not. In the other case (H_C:$\epsilon = 0$), firms in general do not vary prices in the short run.

In actually testing these three hypotheses, it could turn out that none is accepted. This is not just a moot possibility. Whereas there are three coefficients involved and each one could be either less than, equal to, or greater than zero, the three hypotheses account for only one-ninth of the total possible outcomes. Accordingly, if the twenty-seven mutually exclusive outcomes were all equally likely, there would then be almost a 90 percent probability that none of the hypotheses is accepted.

Nor can the "undefined" outcomes be dismissed a priori as being implausible. Indeed, in certain instances, the sign restrictions correspond with those imposed—either explicitly or implicitly—in other empirical pricing models. This can be seen in Table 3. Arrayed there are the partial elasticities from the estimated price equations reported by de Menil, Ben Laden, and Otto Eckstein and Gary Fromm. Among the many differences between these specifications and equation (15), three stand out as being especially pertinent to the problem at hand: 1) Two of the models (like virtually all others) require that prices be *positively* related to the input of labor. Laden's model omits the argument entirely and is unusual in this respect. As a general rule, the man-hours argument is included in the price equation as an element of either unit labor costs or output per man-hour. 2) Two of the models (again like virtually all others) omit the capital-stock arguments. Laden's model is again the exception; but even so, he finds that the contemporaneous value of the capital stock exerts a *positive* effect on prices. 3) The other models all omit the rental price of capital, perhaps because it is not considered to be an element

TABLE 3—ALTERNATIVE MODELS OF PRICE DETERMINATION

	Eckstein and Fromm[a]	Laden[b]	de Menil[c]	Text Equation (15)
Cost Terms				
$\partial P/\partial W$	+	+	+	+
$\partial P/\partial M$	+	+	+	+
$\partial P/\partial R$	...	...	...	+/0
$\partial P/\partial L$	+	...	+	−/0
$\partial P/\partial K$	...	+	...	−/0
Other Terms[d]				
$\partial P/\partial Q$[e]	−	−	−	+
$\partial P/\partial ED$[f]	+	...	+	...
$\partial P/\partial P_{-1}$	+	...	+	...

[a]Eckstein and Fromm, Table 2, p. 1172; equation 2–(6).
[b]Laden, Table 3, p. 88; equation (2).
[c]de Menil, Table 1, p. 134; equation X.
[d]Certain independent variables specific to individual models have been omitted from the listings.
[e]While Q has been defined as real new orders in estimating equation (15), it is defined in real value-added terms in the other models
[f]*ED* denotes excess demand, measured as capacity utilization by Eckstein and Fromm and as the trend-adjusted ratio of unfilled orders to capacity output by de Menil.

of the variable costs of production or else because it did not turn out to be statistically significant.[13] In short, then, it cannot be taken for granted that one of the three hypotheses will be accepted. If it were to turn out that none can be accepted, the validity of the model would then be called into question.

Henri Theil has developed a mixed-estimation procedure, pp. 350–51, that can be used to test the three compound hypotheses. Accordingly, in each case, the equalities and inequalities are regarded as prior information, and Theil's χ^2 test can be applied to determine whether the sample and prior information are mutually compatible. Implementation requires that subjective confidence intervals be formulated for the restrictions. Thus, in the case of inequalities,

[13]For reported but unsuccessful attempts at including the rental price of capital in an estimated price equation, see the studies by Dale Heien and Joel Popkin; de Menil and Jared Enzler.

positive or negative, it was assumed that:

$$|\mu_{\Sigma b_{ij}}| = 0.5 \text{ and } \sigma_{\Sigma b_{ij}} = 0.25$$

What this says, in effect, is that it would be highly unlikely for the absolute value of the nonzero elasticities to exceed one. Similarly, in those cases where the sum of the distributed-lag coefficients is supposed to be equal to zero, it was assumed:

$$\mu_{\Sigma b_{ij}} = 0.0 \text{ and } \sigma_{\Sigma b_{ij}} = 0.05$$

Finally, it was also assumed that the (subjective) covariances between the prior conditions on the sums of the coefficients were all equal to zero.

This prior information was combined with the appropriate elements from the estimated variance-covariance matrix for regression 1-2, and Theil's χ^2 statistic was then calculated for each of the three cases: $\epsilon = 1, 0 < \epsilon < 1$, and $\epsilon = 0$. Listed below are the outcomes of the three separate tests:

Hypothesis	ϵ	χ^2
H_A	$\epsilon = 1$	28.4
H_B	$0 < \epsilon < 1$	6.8
H_C	$\epsilon = 0$	5.0

As there are three degrees of freedom in all three cases, the critical value of χ^2 at the 5 percent level is 7.8. Accordingly, the null hypothesis of compatibility between prior and sample information is rejected in the case of $\epsilon = 1$, but is accepted in the other two cases.

Though indefinite in some respects, the results from this application of Theil's test still provide considerable insight into the pricing process. First and foremost, they can be regarded as a validation of the marginalist approach. For the reasons discussed above, it might well have turned out that all three hypotheses were rejected. This even seemed like the most likely possibility, given the empirical results that had been obtained in past pricing studies (for example, those depicted in Table 3). But, in fact, the hypotheses were not all rejected, and the marginalist model is thereby validated. Second, the results also stand as an outright refutation of the hypothesis of fully flexible prices. In rejecting the hypothesis that $\epsilon = 1$, the inference is that firms do not fully adjust their prices in response to the short-run deviations in marginal costs. Necessarily, then, prices must be regarded as being less than fully flexible in the short run.

It remains an open question, however, whether firms adjust prices in part or not at all to reflect the short-run discrepancies. The estimated coefficients in regression 1-2 are evidently not precise enough to enable Theil's test to discriminate between H_B and H_C. This ambiguity is not surprising, nor are its roots difficult to trace. Within the model, the essential differences between these two hypotheses center on the relationships between output prices and the inputs of both labor and capital. Indeed, prices must either be negatively related or not related at all to *both* variables.[14] Now, look again at the distributed lag coefficients for regression 1-2 in Table 2. On the one hand, all three coefficients on the lagged labor-input variables are significantly less than zero. This is solid evidence supporting H_B, as opposed to H_C. On the other hand, the three coefficients on the lagged capital-input variables are scattered around zero, and only one is significantly less than zero. It is this ambiguity that is reflected in the results of Theil's test. This ambiguity was discussed earlier, and it was suggested that the trouble stems from measurement errors built into the estimates of the capital stock. Until this matter can be cleared up, comparatively little confidence can be placed in the estimated coefficients of the lagged capital-stock variables. (That is, while these results cannot be said to be support for H_B, neither can they be interpreted as contradicting it.)

Thus, although the empirical evidence is inconclusive in a strict sense, it does provide

[14] For the null hypothesis that the coefficients on the lagged labor and capital variables were equal to zero, the calculated value of F was 3.91 in regression 1-2. The null hypothesis would be rejected at the 5 percent level. The trouble with this test, however, is that it does not take into account the sign restrictions on the coefficients.

some basis for tentatively choosing between H_B and H_C. Because of the likelihood of measurement error, the confidence limits on the sum of the capital-stock coefficients ought to be widened in effect. Weighted in this way, the strength of the empirical evidence tilts the balance in favor of H_B—that is, that firms adjust their prices *partially* in response to the temporary deviations between long- and short-run marginal costs. Yet, additional evidence is needed before any final conclusion can be drawn.

IV. Conclusions

This study has examined aggregate pricing behavior in the key *U.S.* nonfood manufacturing sector. The principal objective was to measure the speed of adjustment of prices; this was done by determining the extent to which prices are adjusted in response to temporary differences between short- and long-run marginal costs. The following are the main empirical conclusions: first, the marginalist pricing model itself was formally validated in an econometric testing context. These manufacturing firms do appear, therefore, to behave as expected-profit maximizers. Second, it was found that these firms do not adjust their prices fully in response to temporary differences between short- and long-run marginal costs. In this respect, prices can be said to be less than fully flexible in the short run. (It remains an open question, however, whether these firms adjust their prices in part or not at all to reflect the temporary differences in marginal costs.) Third, many firms, if not all, evidently operate in terms of long-run planning horizons and adjust their prices relatively slowly over time. This provides econometric backing for the recent theoretical macroeconomic analyses that have traced income-constrained multiplier processes to *relatively inflexible* prices.

At the same time, the empirical results of this study open up several new avenues for future research. First, given separate measures of the real stocks of plant and equipment, it would be useful to revise and reestimate the pricing model to take them into account. Perhaps, then, it would be possible to determine whether prices are adjusted in part or not at all in response to the temporary short-run deviations in marginal costs. Second, embedded in the analytical framework of this study is the assumption that the speeds of adjustment of price and quantity are the same throughout all stages of the business cycle. It would be interesting to test this hypothesis empirically. Third, the results indicate that rising interest rates do exert a cost-push effect on prices, via the rental price of capital. This is, of course, a partial effect which operates over the "long haul." Nevertheless, it would be useful to know the extent to which this partial cost-push effect hampers the overall effectiveness of monetary policy in engineering a slowdown in the rate of inflation.

REFERENCES

A. Alchian, "Information Costs, Pricing, and Resource Unemployment," in Edmund S. Phelps et al., eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

K. Arrow, "Toward a Theory of Price Adjustment," in Moses Abramovitz et al., eds., *The Allocation of Economic Resources*, Stanford 1959.

R. J. Barro, "A Theory of Price Adjustment," *Rev. Econ. Stud.*, Jan. 1972, *39*, 12-26.

W. Branson and A. Klevorick, "Money Illusion and the Aggregate Consumption Function," *Amer. Econ. Rev.*, Dec. 1969, *59*, 832-49.

G. de Menil, "Aggregate Price Dynamics," *Rev. Econ. Statist.*, May 1974, *56*, 129-40.

——— **and J. J. Enzler,** "Prices and Wages in the FR-MIT-Penn Model," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1971.

O. Eckstein and G. Fromm, "The Price Equation," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1159-83.

D. Heien and J. Popkin, "Price Determination

and Cost-of-Living Measures in a Disaggregated Model of the U.S. Economy," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1971.

D. Hodges, "A Note on Estimation of Cobb-Douglas and CES Production Function Models," *Econometrica*, Oct. 1969, *37*, 721–25.

B. E. Laden, "Perfect Competition, Average Cost Pricing, and the Price Equation," *Rev. Econ. Statist.*, Feb. 1972, *54*, 84–88.

Axel Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, London 1968.

F. Modigliani and R. Sutch, "Innovations in Interest Rate Policy," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 178–97.

W. D. Nordhaus, "Recent Developments in Price Dynamics," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1971.

J. L. Pierce and J. J. Enzler, "The Effects of External Inflationary Shocks," *Brookings Papers*, Washington 1974, *1*, 13–61.

K. Smith, "The Effect of Uncertainty on Monopoly Price, Capital Stock and the Utilization of Capital," *J. Econ. Theory*, June 1969, *1*, 48–59.

Henri Theil, *Principles of Econometrics*, New York 1971.

J. Tobin, "Keynesian Models of Recession and Depression," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 195–202.

R. Turvey, "Marginal Cost," *Econ. J.*, June 1969, *79*, 282–99.

R. N. Waud, "Man-Hour Behavior in U.S. Manufacturing: A Neoclassical Interpretation," *J. Polit. Econ.*, May/June 1968, *76*, 407–27.

Relative Earnings Mobility in the United States

By BRADLEY R. SCHILLER*

Accumulated evidence suggests that the shapes of the income and earnings distributions in the United States are fairly fixed, perhaps even immune to major changes in economic conditions and public policy (see Edward Budd; Peter Henle; T. Paul Schultz; Lee Soltow; Lester Thurow and Robert Lucas; but contrast Morton Paglin). But we still know very little about how mobile individuals are within that rigid size distribution, despite the abundance of hypotheses that have been offered to "explain" such (im)mobility. On the one hand, individuals may be highly mobile across discrete points of the aggregate distribution, suggesting a conventional game of musical chairs (to the tune of the human capital school fight song) in which the position of the chairs themselves is the only thing that never changes. On the other hand, the rigid shape of the aggregate distribution is equally compatible with a total lack of personal mobility—a game, as it were, that individuals play by remaining in their chairs until the music (played by dual labor market theorists and other structuralists) is over.

These extremes of relative income movement have profoundly different implications for our views of income distribution and economic opportunity, as a number of model builders have stressed. Thus, it is somewhat surprising that so much empirical attention has been devoted to the shape of the size distribution of income or earnings, and so little to the fate of individuals within that distribution.[1] As matters now stand, we have a few clues on the extent of individual status mobility[2] and a lot of theorizing about the shape of individual earnings profiles, but few links between them. The intent of this paper is to help bring these subjects together in an explicit way, while providing new empirical information about individual earnings changes.

The paper begins with a discussion of the relative earnings perspective and the expectations for relative earnings mobility generated by alternative models of labor market behavior. The second section describes the basic data source (Social Security earnings records) and the statistical framework (transition matrices) used for summarizing the empirical observations. This section also summarizes the basic findings and examines their sensitivity to both transitory and cyclical disturbances. These findings are compared with theoretical expectations in Section III, while the fourth section provides a summary and conclusions.

The principal finding of this inquiry is that relative earnings mobility is extensive among employed males, both across and within age cohorts. This finding confirms

*American University. Financial support for this research was provided by the U.S. Employment and Training Administration and the University of Maryland's Computer Science Center. William T. Sutton and David Mercer Wells provided valuable computational assistance, while Henry Aaron, George Akerlof, John Conlisk, Albert Fishlow, Bennett Harrison, and Richard Perlman contributed useful criticism and suggestions.

[1] Even on a purely theoretical level, it is difficult to comprehend how so many theories of aggregate distribution can have been formulated (see Martin Brofenbrenner; Harold Lydall, Jan Pen) without more consideration of how mobile individuals are in terms of relative income and what factors might account for such mobility The recent debate on the appropriate measure of the size distribution of income (see Paglin) typifies the fascination with aggregates. Perhaps Pareto was leading us down the wrong road when he directed us towards universal mathematical characterizations of the income distribution, the kind of inquiries that have encouraged neglect of individual mobility and welfare.

[2] See Peter Blau and Otis Duncan or Herbert Parnes (vol. 1) on the subject of occupational mobility; or Parnes (vols. 2-3), James Morgan, Andrew Kohen, Paul Taubman, and Frank Levy for limited data on income mobility.

the existence of tremendous variation in the shape of individual earnings profiles and thus lends considerable support to those labor market models that predict such variation. Of the models tested, the on-the-job training *(OJT)* variant of human capital models stands out in this regard, while serious doubts emerge about those models that postulate varying degrees of segmentation or immobility. But this "confirmation" of *OJT* models is largely a matter of default, given the poor explanatory power of the alternative models tested. It is also observed that black male workers do not share fully in the general pattern of mobility; these racial differences in earnings mobility provide evidence of at least some selective segmentation (discrimination). Additional findings are highlighted in Section IV.

I. Conceptual Framework

The conceptual approach to the present inquiry is straightforward. If there is considerable variation in the slope of individual earnings profiles, then we may anticipate eventual changes in relative earnings status, or what Jacob Mincer (1974) has called "crossovers." That is to say, we may anticipate that individuals will be mobile across points (ranks) of the earnings distribution. Naturally, alternative specifications of individual earnings functions should lead to different conclusions about the extent of relative earnings mobility. From this perspective, relative earnings mobility provides a simple, if indirect, test of alternative earnings functions.

To focus on relative mobility, we may envision the earnings distribution as a hierarchical ordering of a finite number of ranks. Specifically, consider n ranks in the earnings distribution and the n^2 probabilities $P_{i,j}$ that an individual will move from one rank (i) to another (j) in a given time period. Our initial interest here is to determine the implications of alternative labor market models for the values $P_{i,j}$ in an $n \times n$ transition matrix.

A. *Stratification Models*

The most extreme expectations for a matrix of $P_{i,j}$ can be generated from various models of stratification. Consider, for example, those models that emphasize the importance of parental socioeconomic status in shaping one's own status (for example, Samuel Bowles; the author, 1970). If everyone's opportunities were in fact narrowly circumscribed by parental socioeconomic status (for example, via neighborhood schools and occupational "connections") then we would observe little or no relative status change either across or within generations. Individuals might still experience earnings increases, of course, but only those which were consistent with average growth rates; very little latitude would exist for crossovers. In terms of our conceptual framework, what one would observe in such a world would be a transition matrix equal to the identity matrix (i.e., all $P_{i,i} = 1$ and $P_{i,j} = 0, i \neq j$).[3]

Not all models of stratification are so rigid, of course. Models of discrimination based upon race or sex (rather than class) postulate nondiagonal entries, but expect them to be differentiated by the presence of "preferred" or "nonpreferred" workers (see Gary Becker, George Furstenberg, and Barbara Bergmann). Presumably, what is envisioned is a matrix that confines all nonpreferred workers to the lower ranks of the earnings distribution and thus to the upper-left corner of the matrix bounded by the rank b, where b defines the boundary between preferred and nonpreferred jobs and related earnings. In such a world one would

[3]It is interesting to note that there are other models of labor market behavior that also yield identity matrices. In a world where opportunities were neither constrained nor enhanced by social class factors, natural ability would tend to dominate relative status. From this perspective, relative status depends on genetic rather than social class origins (see Richard Herrnstein, but also John Conlisk). Consequently, observing an identity matrix would not settle the "nature vs. nurture" controversy, but only fan the existing flames.

not anticipate movement across discriminatory boundaries, that is, one would expect all $P_{i<b,j>b} = 0$ and $P_{i>b,j<b} = 0$, although there might be considerable latitude for mobility within the submatrices defined by such boundaries.

The more general model of dual labor markets yields a matrix in which mobility is similarly bounded, but not exclusively by race or sex. The essential feature of such models is the duality barrier that separates "primary" (or core) from "secondary" (or peripheral) markets. It is asserted that few workers hurdle this barrier (see Peter Doeringer and Michael Piore; David Gordon; Michael Wachter). Although the full character of neither secondary nor primary jobs has been spelled out, the suggestion has been made that annual earnings are a reliable guide for distinguishing among the two kinds of jobs (see Howard Wachtel and Charles Betsey). From this perspective, then, one can again envision a borderline rank D which reliably separates primary from secondary workers, and thus submatrices of $P_{i,j} = 0$.

Dual labor market theorists are not so explicit, however, about their expectations for the remaining cells. One might suggest, though, that mobility between ranks *within* the secondary market is essentially random, since it is usually postulated that age and experience do not pay off in that market. That is to say, there are no "better" or "worse" career paths in the secondary market that will systematically alter relative earnings positions. But what about earnings mobility within the *primary* market? Piore (1973, 1974) has suggested that relative wages in the primary sector are rigidly fixed by custom, implying very little latitude for changes in relative earnings position, especially when adjusted for years of work experience (or age cohort). This suggests a transition matrix in which the primary market has the character $P_{j,k=j>d} = 1$, with zeros in the remaining cells of the submatrix. Thus, the theoretical expectations generated by these specific versions of the dual labor market model can be summarized as:

$$P_{j,k} = 1/d \text{ for } j, k < d$$
$$P_{j,k} = 1 \text{ for } j = k > d$$
$$P_{j,k} = 0 \text{ for } j > d, k < j \text{ and } j < k, k > d$$

B. *Human Capital Models*

The kinds of mobility expectations generated by models of labor market segmentation can be distinguished readily from those generated by human capital models. As suggested earlier, the basic message of human capital theories is that individuals possess the power to alter their lifetime stream of earnings by making alternative sacrifices and investment decisions (see Becker; Lee Lillard; Mincer 1970, 1974). If this is the case, then one would anticipate that individuals experience distinctly different earnings streams over time. In particular, if human capital theorists are correct in assuming that investment in one's later earnings potential entails a sacrifice of present earnings, then one might expect to observe considerable mobility in relative earnings ranks as individuals experience the burdens and payoffs of their varying investment decisions.

Our expectations vis-à-vis relative earnings mobility are sensitive, of course, to the kinds of investment we think important. If all human capital investment took place prior to labor market entry, then one would observe few changes in relative earnings positions once everyone had entered the labor market. In effect, everyone would be assigned a permanent position (rank) in the earnings distribution on the basis of the human capital they brought to the labor market. As Taubman has argued, even if that capital is not immediately observable at the time of entry, employers will soon differentiate among workers on the basis of performance. Under these conditions, workers will move quickly into permanent relative positions (all $P_{j,j} = 1$), although the dollar distance between those positions might grow over time.

A very different set of expectations is generated by those human capital models that emphasize on-the-job training. In par-

ticular, one should expect to observe more mobility between discrete points of the earnings distribution (what Mincer, 1974, calls crossovers and I observe as $P_{i,j} \neq 0$) where human capital investments are assumed to take place in the labor market itself; that is, where experience (on-the-job training) is an important determinant of the slope of individual earnings functions. In view of the increasing recognition given to on-the-job training and investment (see William Haley; Thomas Johnson and Frederick Hebein; Edward Lazear; Mincer, 1974; Sherwin Rosen), it seems reasonable to anticipate considerable crossing of relative earnings positions, at least from this particular view of labor market dynamics.

The difficulty with the *OJT* variant of human capital models is the empirical need to specify what is meant by a "considerable" amount of crossovers. One could argue that the opportunities for on-the-job training are so numerous that all individuals have the chance to move to any point in the distribution. From this perspective, one might anticipate a matrix with all $P_{i,j} = 1/n$, where the relative position of each individual is determined by his tastes, his discount rate, and the duration of the relevant investment and payoff periods.

Although the expectation of all $P_{i,j} = 1/n$ is not inconsistent with the *OJT* model, it does not fully reflect the richness of the model. In particular, Mincer and others argue that the amount of *OJT* investment is reflected in the difference between actual and potential (opportunity) earnings during the investment period, with the payoff expressed as the excess of later earnings over what they would have been in the absence of such investment. In terms of our conceptual framework, this implies that individuals of given ability who begin in the lower ranks of the earnings distribution should experience more upward mobility than others of equal ability but higher initial earnings positions. To test this hypothesis, one would have to identify workers of "equal ability," something we are unable to do with the data available. Within the confines of the present data set, we could test this hypothesis only if we were willing to assume that all workers entered the labor market with identical potential (ability) and also that our observation period (15 years) captured the bulk of the investment and payoff period, in which case one could postulate a transition matrix with all minor diagonal elements equal to one, that is, $P_{j,n-j} = 1$. Neither assumption is very palatable. Accordingly, we are constrained to evaluate the *OJT* model only in the equiprobabilistic form described above.

C. *Other Models*

A perspective which combines some of the features of both labor market segmentation and human capital theory has been dubbed the job competition model (see Thurow). According to this view, marginal productivities are inherent in jobs, not people. Thus workers compete for access to a fixed distribution of marginal productivities (jobs), either on the basis of their trainability or employer prejudices. Relative earnings positions are then determined by the outcomes of the job competition. An interesting (and testable) implication of this particular view is that movements between discrete points of the earnings distribution will tend to be accompanied by job changes, either across firms or within the bureaucratic structure of large firms (see David Wise).[4]

Finally, we may note that many observers regard the outcomes of the labor market as essentially random, being attributable to the vagaries of fortune (see Christopher Jencks), perhaps conditioned by the willingness to incur risk (see Milton Friedman). Such views of the labor market obviously lead to the expectation of high rate of relative earnings mobility. In particular, if initial and final ranks were in fact statistically independent, one would anticipate all $P_{i,j} =$

[4]The significance of job changing for earnings growth has also been emphasized by George Borjas and Ann Bartel, who stress the implications for loss of *OJT* training, and by Duane Leigh, who emphasizes occupational change as a determinant of earnings growth.

$1/n$, an expectation that bears a notable resemblance to the *OJT* human capital model discussed above.

Other stochastic processes are possible, of course, and can in fact be contrived to generate just about any transition matrix desired. Perhaps of greatest interest are those models which explicitly incorporate an element of luck (randomness) into models of opportunity stratification, thereby creating distinct mobility tracks, but making the assignment of individuals to such tracks an essentially random process (see Wise). Such models focus on the *conditional* probability of mobility, suggesting that people who have started moving—having been (randomly) selected for a mobility track—will tend to continue moving. These models can generate transition matrices identical to those of *OJT* models, despite their different implications regarding the role of individual decision making in the mobility process.

One may safely assert, then, that competing models of labor market behavior do imply different relative earnings patterns over time. They are not so well-specified, of course, that one can expect to identify a given amount of mobility as uniquely verifying a particular model; ultimately, one can distinguish between them only on the basis of subjective judgements about what constitutes a little or a lot of relative earnings mobility. In so doing, however, we may provide some important perspectives on labor market behavior and related issues of income distribution.

II. The Data

In order to determine the extent of mobility in relative earnings over time, we need to know the actual distribution of earnings in specific years, as well as each worker's position therein. For this purpose we employ the Longitudinal Employer Employee Data (*LEED*) file of Social Security Administration (*SSA*) records, which contains quarterly observations on individual earnings histories for 1 percent of all covered earners. The unique and decisive advantages of the *LEED* file are 1) sheer volume, 2) longitudinal continuity, 3) nearly universal coverage (over 90 percent of all wage and salary workers are now covered), 4) detail on firm and industry attachment, and 5) reliability. On the last point it is important to note that approximately 20 percent of all workers have earnings in excess of *SSA* tax ceilings, and that *SSA* extrapolates from quarterly earnings to derive annual earnings estimates for this group. Although these estimates turn out to be very crude approximations for the highest income groups, they are adequate for assigning individuals to broad subdivisions of the aggregate distribution. What makes the *SSA* data particularly appealing from a theoretical point of view is that they focus on labor earnings alone, and thus on the outcomes of labor market processes.[5]

Although the *LEED* file is obviously well-suited for an inquiry into earnings mobility, it is not perfect. Of particular concern is the absence of data on education and occupation, which limits our ability to explain the mobility patterns we observe. In addition, *SSA* records cannot distinguish between a move from covered employment to noncovered employment (principally federal and various state and local jobs), and a move to unemployment or nonparticipation status. Hence, the data is best-suited for a study of mobility among persons continuously in covered employment, and we shall concentrate on this subpopulation.[6]

The present inquiry focuses on a sample of 74,227 males from the *LEED* file. To be included in the sample, male workers had to satisfy the following conditions:

(i) between the ages of 16 and 49 in

[5]See Thurow for a discussion of the differences in labor and nonlabor income determination. We should also note that annual earnings are the largest component of total income as well as the largest source of variability in family incomes over time (see Morgan, James Smith).

[6]Not all covered earnings are reported, of course; many employers (most notably employers of domestic help) do not report wages or pay the required taxes. But this problem is not likely to affect our results significantly, as we restrict our observations to male workers with substantial work experience.

1957, the first year of our observation period;
(ii) at least $1,000 of earnings in 1957;
(iii) earnings in 1971, the final year of our observation period.

Conditions (i) and (iii) are imposed to assure a sufficiently extended longitudinal framework, without including workers who begin working at unusually early ages or continue working past typical retirement ages. Condition (ii) is imposed to exclude the "mobility" of younger workers moving from part-time to full-time work.[7] The objective of the inquiry is to determine how the earnings of those workers who meet the above conditions ("attached workers") changed during the subsequent 14 years, relative to other male workers, that is, the extent to which these workers moved from one rank of the earnings distribution to another.

A. *The Statistical Framework*

The empirical work reported here focuses on cohort-specific mobility, rather than mobility within the entire earnings distribution.[8] Thus, we focus on the question of whether or not individuals of approximately the same age (and presumably experience) exchange relative earnings positions over time, ignoring changes in relative status brought about by differences in age. This cohort-specific focus is desirable to the extent that individuals gauge their status in relation only to others of the same age (and sex and race?)[9] and to the extent that the various labor market models discussed here emphasize the qualitative content of labor market experience.

To isolate intracohort movement, I have reconstructed the earnings distributions specific to each five-year age cohort, in every year of the observation period, using the *entire* 1 percent *LEED* file of male workers as a data base. Each year's cohort-specific distributions have then been subdivided into twenty proportional, hierarchical ranks or "ventiles."[10] With this information we can assign each of our 74,227 sampled male workers to a ventile rank in the earnings distribution for his cohort. The question of "relative earnings mobility" may then be addressed by determining whether or not an individual moves to another cohort-specific ventile ranking in subsequent years. The resultant moves comprise our empirical estimates of $P_{i,j}$.[11]

Table 1 provides summary measures of the $P_{i,j}$ calculated for the period 1957–71.[12] There are no observations in the first two rows of Table 1 since workers who earned less than $1,000 in 1957, enough to exceed the first two ventile ceilings in all 1957 cohorts, were excluded (thus, $P_{1,j} = P_{2,j} = 0$). Accordingly, our observations begin with individuals who were in the third cohort ventile position (*CVP*) in 1957, generally a group of very low-income workers.

We may begin to gauge the extent of relative earnings mobility by noting that the

[7] Out of a total available sample of 88,203 men, 13,976 were excluded from consideration by the earnings cutoff. Nearly half of these were under the age of 20, and nearly 80 percent were under 30. Subsequent analysis of these excluded individuals suggested that their mobility patterns were not strikingly different from those described below.

[8] It is worth noting that the intracohort mobility we observe here accounts for most of the mobility of our sample; that is to say, intracohort mobility overwhelms intercohort mobility—experience per se is not as important as the nature of that experience in determining relative earnings growth. Additional tabulations on intercohort mobility are available from the author.

[9] This is implicit in Paglin.

[10] John McCall also partitioned the earnings distribution into "relative earnings" ranks, but his partitions were based on arbitrary fractions of median earnings, and thus not proportional in size.

[11] Note that our relative earnings approach automatically adjusts for changes in money wages (inflation) and for any shifts that might be occuring in the aggregate (size) distribution. What it does not tell us, of course, is whether or not the mean distance (measured in dollars) between points of the distribution is increasing or not, a measure of change in the size distribution of earnings and hence of equality in status. More specific consideration of these issues is contained in McCall, and Nancy Ruggles and Richard Ruggles; both use Social Security records as an empirical base.

[12] A complete matrix of $P_{i,j}$ as well as the specification of ventile boundaries is available from the author.

TABLE 1—SUMMARY MEASURES OF INTRACOHORT MOBILITY, 1957–71

1957 Cohort Ventile	Mean Absolute Change	Percent Immobile	Mean Algebraic Change
1	–	–	–
2	–	–	–
3	4.26	35	3.57
4	4.01	33	2.90
5	4.04	30	2.56
6	4.31	26	2.54
7	4.08	26	1.91
8	4.13	25	1.33
9	3.98	24	.64
10	4.05	23	.31
11	3.82	24	– .54
12	3.92	24	– .87
13	3.93	24	–1.48
14	4.27	25	–2.32
15	4.34	25	–2.82
16	4.45	27	–3.30
17	4.56	28	–3.73
18	4.75	31	–4.27
19	4.51	42	–4.19
20	4.12	48	–4.12
Total	4.22	29	– .93
Standard Deviation	(3.82)		(5.61)

Note: Correlation coefficient = .150; N = 74,227.

correlation coefficient for this sample is .15, suggesting that there is virtually no linear relationship between $CVP_{i(1957)}$ and $CVP_{i(1971)}$. The other measures of mobility depicted in Table 1 attempt to convey the extent of movement experienced by this highly mobile sample. The mean absolute change (average absolute deviation) is 4.22 ventiles (approximately 21 percentiles). The extent of mobility experienced by individuals from each 1957 ventile is shown as well; the figures suggest that the extent of mobility experienced does not vary substantially across 1957 ventiles.

Mean absolute changes in relative position may, of course, disguise a lot of immobility if the means are overly influenced by the experiences of a small number of highly mobile people. Accordingly, a second measure of mobility is provided in Table 1, namely, the percentage of people who move less than two ventiles. A move from one ventile to another may encompass a distance of anywhere from 0 to 10 percentiles and may thus obscure a lot of immobility. Hence, we designate $\Delta CVP < 2$ as "little or no mobility." By this standard, 29 percent of the entire sample was immobile, although there are marked differences across 1957 ventiles. Notice in particular the significantly higher rates of immobility in the highest and lowest 1957 ventiles: the relative status of people at the top or bottom of the earnings distribution is significantly more stable than the status of those in the broad middle ranges of the distribution.[13]

Finally, we present the mean algebraic movement of the sample and each of the 1957 cohort ventiles. The overall algebraic decline in status results from the imposition of a floor under 1957, but not 1971 ventile positions; otherwise the mean algebraic move for the entire sample would be zero and of no interest. In examining the algebraic movement of the separate 1957 ventiles, a "regression towards the mean" is clearly discernible, implying more equality of lifetime earnings than is evident in the figures for any particular year.

B. *Transitory and Cyclical Disturbances*

The degree of significance to be attached to observed mobility patterns depends in part upon how "permanent" such patterns are thought to be. On one hand, there is a distinct possibility that much or all of the observed mobility is of a transitory nature, and thus of little long-run statistical or socioeconomic significance (see Friedman; Jencks). On the other hand, there is also the very real possibility that the relative earnings mobility observed here is unduly influenced by cyclical factors. This second concern is accentuated by the fact that the *U.S.* economy experienced a moderate recession near the end of the observation period (1971).

[13]The same pattern of immobility across 1957 ventiles was observed using $\Delta CVP \leq 3$ as the standard for "immobile," although the average level of immobility (42.4 percent) was of course higher.

In examining the question of whether transitory or cyclical phenomena have distorted our perceptions of mobility, it must be remembered that we are measuring relative earnings mobility across *ventile* positions of the earnings distributions. Furthermore, mobility was defined as a move of at least *two* ventiles (i.e., a move of from 5 to 10 percentiles). Accordingly, the basic definition of immobility used here allows for considerable variation in relative incomes, be it of a random or cyclical nature. That is to say, our statistical framework establishes the *presumption* that our mobility observations represent more than random or cyclical noise.[14] This presumption does not obviate the necessity to test for transitory and cyclical disturbances, however.

The test for "transitory" noise takes advantage of the longitudinal character of the data set. The data are first partitioned into five-year subperiods, and within each period individuals are classified according to their mobility experience during that period. The question thus becomes whether or not the mobility of one period is reversed in the next, thus rendering mobility a transitory phenomenon.[15]

The focus here will be on the periods 1957–62 and 1962–67.[16] Individuals are first classified according to their observed (im)-mobility during the period 1957–62, then the mean change in *CVP* is computed for each of these groups over the period 1962–67. Thus, we are implicitly attempting to predict 1962–67 mobility on the basis of 1957-62 experience. For convenience of exposition, the test statistic Q is computed according to

$$Q = \frac{\bar{\Delta} CVP_{1967/1962}}{\bar{\Delta} CVP_{1962/1957}}$$

The sign of Q indicates whether or not the direction of mobility has been changed, while the absolute value of Q gives the extent to which the average rate of mobility has been altered. Clearly, a Q-value of -1.0 or smaller suggests that observed mobility is "transitory" in the sense of being reversed in subsequent years. On the other hand, positive Q-values indicate constant direction of movement, and thus suggest that mobility is understated by the above measurements. Finally, Q-values between 0 and -1.0 indicate some degree of reverse mobility, but not enough to offset earlier gains or losses. The Q-statistics have been calculated for the downward and upward individuals within each 1962 *CVP*, and are presented in the first two columns of Table 2. Empty cells reflect our 1957 income floor and the underlying definition of mobility ($\bar{\Delta} CVP \geq 2$).

A review of the first two columns suggests that full reversal of observed 1957–62 mobility occurs in only 1 out of 34 possible cells, and this at one of the extreme ends of the distribution (where floor and ceiling ef-

TABLE 2—Q-VALUES, BY 1962 *CVP* AND MOBILITY EXPERIENCES, 1957–62[a]

1962 Cohort Ventile	1957–62 Mobility Experiences Up	Down
1	–	–1.03
2	–	– .82
3	–	– .74
4	–	– .64
5	.30	– .52
6	.06	– .46
7	–.15	– .35
8	.27	– .39
9	.24	– .38
10	.04	– .22
11	.00	– .09
12	–.07	.04
13	–.16	.11
14	–.28	.30
15	–.33	.33
16	–.50	.70
17	–.54	.61
18	–.49	.34
19	–.44	–
20	–.53	–
Total	–.27	–.20

[a] $Q = \frac{\bar{\Delta} CVP_{(i)1967/62}}{\bar{\Delta} CVP_{(i)1962/57}}$

[14] At the same time, of course, this framework makes it more difficult to ascertain the full significance of transitory income changes, especially those of an intra-ventile nature.

[15] For alternative tests see Lillard, and Lillard and Yoram Weiss.

[16] Similar analyses were done for other subperiods and yielded analogous results.

fects are most important). Further inspection suggests only a few Q-values are close to -1.0 and thus of support to the "reversal" hypothesis. The remaining cells, of course, reflect either partial retrenchment or constant direction of mobility. Taken together, these observations suggest that the reversal hypothesis has some validity, but also that most of the mobility observed here is in fact "permanent." Without observing the entire work cycle, one cannot of course preclude the possibility of later reversals. But these findings are strong enough to reject the suggestion that our present observations reflect nothing more than transitory disturbance.[17]

Additional evidence on transitory disturbances can be generated by varying the length of the observation period. If transitory noise dominates our observations, then the above mobility measures should not be sensitive to the number of years that elapse between initial and terminal observations of relative earnings. On the other hand, if there really are mobility *patterns*, that is, individuals trending up or down the distribution, then measured mobility should increase with the length of the observation period. This simple test of transitory disturbances is accomplished by shortening our observation period to ten years, that is, to the period 1957-67 from 1957-71. This truncation of our longitudinal horizons does in fact significantly reduce measured mobility. Specifically, the correlation coefficient increases from .150 to .187, the mean absolute move decreases from 4.22 to 3.54 ventiles, and the proportion of immobile workers increases from 29 to 34 percent. Moreover, the decrease in mobility is experienced by all ventiles and age cohorts.

The increase in mobility that results from adding four years to the 1957-67 period may be due to the cyclical factors mentioned above, however. To test this hypothesis one can compare mobility rates for the periods 1957-67 and 1962-71.[18] If the latter period evidences significantly more mobility than the former, then the possibility of cyclical distortions should be taken seriously. This is not the case. The correlation coefficient for the 1962-71 period is .192 (vs. .187), the mean absolute move is 3.67 (vs. 3.54) and the proportion immobile is 35 percent (vs. 34 percent). Accordingly, there is no evidence that the above measures of mobility are seriously inflated by cyclical factors unique to the terminal year chosen.

It appears, then, that our basic empirical observations are sound, in the sense of reflecting more than transitory or cyclical disturbances. In the following section I shall attempt to assess the implications of such mobility for alternative labor market models.

III. Tests of Expectations

As interesting as the basic findings may be in their own right, it is still difficult to determine whether they reveal a lot of mobility or just a little. How significant is it that 70 percent of male workers are mobile, and that the average move spans a distance of one-fifth of the aggregate distribution? As a partial answer to this question, one can compare our findings to the expectations previously generated from alternative labor market models. The summary measures depicted in Table 1 serve as the foundation for such tests, although submatrices and additional summary measures are generated for blacks and for different age cohorts as the testing proceeds.

A. *Stratification Models*

1. *Class Discrimination*

As noted earlier, models of rigid segmentation based on class or parental socioeconomic status imply little or no movement off the major diagonal (i.e., $P_{i,j}$ =

[17]It should be noted that we have not said anything here about the *process* which generates (permanent) mobility. It is quite possible, as we shall note again in Section IV, to generate our observed mobility patterns from stochastic processes of (re)distribution.

[18]The use of 1962 rather than 1961 is dictated by our early formation of the basic data file into five-year observation periods; this does not influence our conclusions here, however, as we shall note.

$0_{,i \neq j}$). Such models are clearly incompatible with observed patterns. Only 10.6 percent of the workers in this sample satisfy this expectation, the other 89.4 percent representing deviations. Even on the basis of the more liberal measure of immobility $\Delta CVP < 2$, it was found (Table 1) that only 29 percent of the sample lived up to the expectations generated by this model. Finally, recall that the correlation coefficient between CVP_{1957} and CVP_{1971} amounts to only .15, suggesting pervasive mobility.

What little support the class discrimination model does find in this data lies at the extremes of the earnings distribution: over 33 percent of the lowest paid (attached) workers remain at the bottom of the distribution, while well over 40 percent of those in the highest reaches of the distribution remain in their position over the fifteen years of the observation period. Even this support must be qualified, however, by the open-ended nature of the highest earnings rank: once individuals exceed the lower boundary of the highest ventile they may experience tremendous changes in earnings but still be counted as immobile.

2. *Racial Discrimination*

The failure of the most extreme segmentation model to account for observed mobility patterns does not deny the possibility of discrimination against particular subpopulations of any class, of course. Accordingly, an explicit test of the racial discrimination hypothesis will be considered next. To do so, a transition matrix was constructed for the 6,109 blacks in the sample, with their relative status determined by their earnings in comparison to all male workers of the same age cohort.[19]

The simplest version of the discrimination model, namely, the notion the blacks are confined to (or "crowded into") the lower end of the earnings distribution is not supported by our observations. To test the model, the hypothetical boundary b, which separates blacks and whites, must be identified. Without stipulating b a priori, we can look instead for such a demarcation in the data itself. What we find is that the boundary, if it exists, is high up into the distribution. Only 17 percent of the blacks in our sample are confined to the first five ventiles, 50 percent to the first nine, and we must go as far as the twelfth ventile to capture 70 percent of the blacks.

This is not to deny any evidence of discrimination, of course, but only to reject the crudest models of it. There are alternative formulations. One might hypothesize, for example, that individual blacks enter the labor market with different bundles of skills and thus start out in a variety of ventile positions. Having begun their careers, however, they confront restricted opportunities for further mobility. This hypothesis is compatible with the above observation that there is no meaningful boundary between blacks and whites on the earnings scale; what it suggests is that blacks will experience less mobility than whites once their starting position is determined.

Our summary measures of mobility provide mixed evidence on the comparative mobility of blacks. Black males are on average no more immobile than white males (30 vs. 29 percent), according to our basic standard of immobility ($\Delta CVP < 2$). But blacks experience a mean absolute move of only 3.73 ventiles compared to average white mobility of 4.22 ventiles, a difference which is both statistically significant (at the .005 level) and important. The most striking differences between blacks and whites, however, are apparent when the patterns of mobility across ventiles are examined, as in Table 3. Notice in particular the extremely low rates of immobility among blacks at the top of the distribution (for example, $CVP = 19, 20$), and their very high rate of immobility at the bottom. What this means is that black workers have an easier time staying at the bottom of the distribution, but a difficult time precariously clinging to the higher earnings positions.

Our final measure of racial discrimina-

[19]Recall that the earnings distributions themselves are generated from the entire *LEED* file. This and subsequent matrices are available on request from the author.

TABLE 3—COMPARATIVE BLACK MOBILITY, 1957-71

1957 Cohort Ventile	Percent Immobile		Mean Algebraic Change	
	White	Black	White	Black
1	-	-	-	-
2	-	-	-	-
3	35	47	3.57	1.93
4	33	44	2.90	1.39
5	30	39	2.56	.88
6	26	37	2.54	.58
7	26	29	1.91	.42
8	25	26	1.33	- .36
9	24	20	.64	- .67
10	23	23	.31	- .55
11	24	22	- .54	-1.76
12	24	26	- .87	-2.34
13	24	23	-1.48	-2.20
14	25	20	-2.32	-4.44
15	25	17	-2.82	-5.15
16	27	15	-3.30	-6.10
17	28	15	-3.73	-6.83
18	31	16	-4.27	-7.68
19	42	10	-4.19	-7.97
20	48	07	-4.12	-7.21
Average	29	30	- .93	- .77

tion is the mean algebraic change in relative earnings positions experienced by whites and blacks. The concern here is to determine whether or not blacks and whites who begin in the same 1957 position move equal distances therefrom. The above observations on immobility rates already suggest a negative answer, of course, but the algebraic deviations provide the clearest picture of just how difficult it is for blacks to hold onto high relative earnings positions. Notice in Table 3, for example, that blacks from the highest 1957 ventiles fell on average over three ventiles further than similarly positioned whites. On the other hand, blacks who started out in the lower ventiles failed to match the high upward mobility of whites from those same ranks.[20]

As a concluding observation, we may note that, overall, black workers failed to increase their relative status over the period 1957 to 1971. Their relative gain of .15 ventiles is not only inconsequential from a socioeconomic perspective, but even fails to achieve statistical significance at the .01 level, something quite unusual for this sample. What this observation suggests is that the civil rights and equal opportunity iniviatives of the 1960's failed to benefit black workers who were already assimilated into the labor market. At best, it appears that such activity benefited only black entrants into the labor force, workers who would not be included in our sample of attached workers.[21]

3. *Dual Labor Markets*

As was noted in Section I, the dual labor market model cannot be meaningfully tested unless one is willing to identify the location of the boundary *d* between primary

[20]All of the ventile-specific differences between the algebraic mobility of blacks and whites as reported in Table 3 are statistically significant at the .001 level. McCall performed a similar analysis, calculating expected absolute incomes in 1966 for blacks and whites of similar 1957 incomes.

[21]Victor Fuchs has suggested that this conclusion is not fully warranted, as the relative status of blacks might actually have fallen in the absence of the civil rights movement; but this is a very limited concept of success (and itself unproven).

and secondary markets, and formulate explicit hypotheses about relative earnings behavior in each market. I have chosen to locate d at the fifth ventile, and compared the transition matrix of observed $P_{i,j}$ to the dual labor market model on this basis. As it turns out, the substance of our findings is not sensitive to this choice. The dual labor market model does little better than the more rigid class discrimination model in accounting for observed mobility. Overall, only 14.8 percent of the sample fulfills the expectations of either remaining in the secondary market or in a fixed relative position within the primary market. Using the broader measure of immobility ($\Delta CVP < 2$) in the primary market, the predictive capability of the model increases to 30.5 percent of the sample. This is still quite modest in view of the fact that a perfectly random distribution of people across ventiles would "explain" half of that.

It could be argued, of course, that our sample of attached workers does not really represent the population envisioned by dual market theorists, particularly with respect to the secondary market. It is often suggested, for example, that women, blacks, and teenagers comprise a substantial proportion of the secondary labor market. Hence, it might be argued that one cannot really disprove the duality hypothesis by observing high rates of mobility among a general sample of attached male workers.

The objection is important, but not entirely convincing. Our data at least suggest that vast numbers of males move out of low paying jobs into better ones, thus refuting the notion of a self-contained trap. We have also demonstrated that the simple duality model does not apply to black males. As for teenagers, the model can only suggest that the kinds of jobs available to young inexperienced labor market entrants are limited, not that people who begin work at young ages never climb the relative status ladder.[22]

[22]The workers in our sample who were aged 16–19 in 1957 experienced extremely high upward mobility over the ensuing fourteen years, rising an average of 9.1 ventiles in the aggregate (not cohort-specific) earnings distribution.

Accordingly, to the extent that earnings are a key descriptor of primary or secondary jobs, one may conclude that the duality model is not consistent with observed mobility patterns.

B. *Human Capital Models*

1. *Schooling Investments*

In testing human capital models one should maintain the distinction between schooling and on-the-job training investments. As was observed earlier, the basic schooling models generate (im)mobility expectations similar to those of class discrimination models; thus, they fail to account for the mobility patterns documented here. Even Taubman's suggestion that it may take a few years for workers to find their appropriate human capital slots does not vindicate the schooling model. This can be seen in Table 4, which depicts the mobility experiences unique to each cohort. According to Taubman's argument, very little mobility should be experienced by workers over the age of 25. Examination of Table 4 clearly indicates, however, that mobility is a pervasive phenomenon for all age cohorts, despite the fact that rates of mobility (as measured by percent immobile or mean absolute change) tend to decline with age.

2. *OJT Investments*

As was noted in Section I, the expectations of *OJT* models are difficult to specify, particularly when the length of the work cycle is truncated and there is no independent control on an individual's initial stock of human capital. It was also noted, however, that an expectation of $P_{i,j} = 1/n$ (or .05 in our framework) is not unreasonable if *OJT* opportunities are pervasive. Using this expectation as a test, it was found that the *OJT* model performs quite well, correctly anticipating over half (54.8 percent) of our observations. If we restrict our observations to the broad middle range of the distribution, cutting off the highest and lowest two ventiles, the goodness of fit rises to 60.5 percent. On this basis then, it appears that the *OJT* model of human cap-

TABLE 4—INTRACOHORT MOBILITY MEASURES, BY COHORT AND RACE

1957 Cohort	Mean Algebraic Change		Mean Absolute Change		Percent Immobile		N
	Total	Black	Total	Black	Total	Black	
16–19	−5.60	−8.79	6.49	9.09	.17	.06	3547 total
	(5.72)	(5.10)	(4.68)	(4.55)			198 blacks
20–24	− .56	−2.03	5.23	4.43	.20	.23	11839
	(6.50)	(5.16)	(3.89)	(3.33)			1064
25–29	− .70	− .74	4.46	3.82	.24	.26	13401
	(5.81)	(4.87)	(3.79)	(3.11)			1152
30–34	− .48	− .00	3.79	3.41	.31	.33	13509
	(5.18)	(4.60)	(3.56)	(3.09)			1130
35–39	− .74	− .16	3.68	3.08	.33	.35	12672
	(5.06)	(4.13)	(3.55)	(2 76)			990
40–44	− .85	.07	3.62	3.08	.35	.37	11028
	(5.04)	(4.22)	(3.60)	(2.88)			922
45–49	−1.00	13	3.72	3.28	.35	.35	8231
	(5.18)	(4 46)	(3.74)	(3.02)			670

Note: Standard deviations in parentheses

ital development derives considerable support from the data.

C. *Other Models*

The high rates of relative earnings mobility observed here are better accounted for by *OJT* models of human capital investment than the stratification or education models reviewed. However, the tenuous nature of that support is apparent when we reconsider other models, particularly those based on some form of stochastic mechanism for distributing relative status.

As noted in Section I, a completely random redistribution of status in every year would generate transition probabilities of $P_{i,j}$ = .05. In this sense, it would appear that a random distribution process explains our observations just as well as the *OJT* model. However, we have also noted a semblance of stability in the direction of mobility, as detailed in Table 3. Since a completely random (re)distribution process would not generate directional stability, it does not explain our observations as well as the *OJT* model of human capital investment.

A completely random redistribution in every year is an extreme version of the stochastic process, of course, and is theoretically unsatisfying as well. More satisfying are models which base the stochastic process on some form of substantive foundation, be it personal characteristics (see Bartel) or shifts in the structure of opportunities (see Aage Sorenson; Wise; Thurow). Furthermore, limited attempts to compare the expected outcomes of such processes with empirical observations have been partially successful (see Wise; McCall). What must be emphasized here is that a stochastic distribution process could be specified that would yield good "predictions" of the mobility observed here, including the element of directional stability. (Wise is particularly interesting in this regard.)

IV. Summary and Conclusions

The longitudinal earnings data reviewed in this paper unambiguously demonstrate that individuals are highly mobile across relative positions in the earnings distribution: 70 percent of the male workers in our sample are mobile and move on average a distance of one-fifth across the earnings distribution of their age cohort. Moreover, high mobility is a phenomenon that continues throughout one's working life. It wa

also noted that an element of directional stability exists in this mobility, implying that these observations reflect something more than purely random fluctuations.

Models of labor market behavior that imply high rates of individual mobility will naturally find the greatest support in our observations. Of the models reviewed, the *OJT* variant of human capital models stands out in this regard, explaining more than half of the observed mobility. However, because other models (particularly those that distribute relative positions on the basis on some stochastic process) could generate equally good predictions, this support must be severely qualified. Essentially, what we have demonstrated is that those labor market models implying high rates of relative earnings mobility have more *ex post* plausability than various models of stratification. The *OJT* model is not uniquely verified by such observations; rather it stands out as one of the few models that formally accounts for mobility.

Our conclusions must be further qualified by the pockets of immobility observed. Although, for example, it is clear that models of pervasive stratification (on whatever basis) are not likely candidates for *general* descriptions of the labor market, they may still play important roles. The lower mobility of black workers, for example, is obviously consistent with such models, as is the tendency toward lower mobility in the highest and lowest ventiles. The evidence on blacks is particularly disturbing as it not only suggests differential constraints on mobility but also that black workers already assimilated into the labor market by 1957 failed to receive any relative benefits over the subsequent fourteen years, a period which encompassed extensive civil rights and equal opportunity activity.

The concept of relative earnings mobility provides a new perspective for assessing labor market models and behavior, as well as for interpreting the microeconomic substance of the (static) distribution of income. It cannot yet answer all the questions one might like to ask, however. To distinguish still more reliably among alternative earnings models, one must answer additional research questions. In particular, the relationship between earnings mobility and job experience must be examined. A most compelling implication of these findings is that more attention should be focused on the *process* of mobility, that is, on the mechanisms that facilitate or obstruct status improvement. The *OJT* model is but one example of such a focus, and has come to dominate this perspective, largely by default. Alternative explanations could easily be formulated, either on the basis of other supply characteristics, demand-side variables, or institutional factors which alter the character of both supply and demand. Presumably, the failure to recognize the pervasiveness of relative earnings mobility has inhibited the development of such alternatives.

REFERENCES

A. J. Alexander, "Income, Experience, and the Structure of Internal Markets," *Quart. J. Econ.*, Feb. 1974, *56*, 63–85.

A. P. Bartel, "Job Mobility and Earnings Growth," Nat. Bur. Econ. Res. work. paper no. 117, New York 1975.

Gary Becker, *The Economics of Discrimination*, Chicago 1971.

———, *Human Capital*, New York 1964.

B. R. Bergmann, "The Effect on White Incomes of Discrimination in Employment," *J. Polit. Econ.*, Mar./Apr. 1971, *79*, 294–313.

Peter M. Blau and Otis Dudley Duncan, *The American Occupational Structure*, New York 1967.

G. Borjas, "Job Investment, Labor Mobility and Earnings," unpublished doctoral dissertation, Columbia Univ. 1975.

S. Bowles, "Understanding Unequal Economic Opportunity," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 346–56.

Martin Brofenbrenner, *Income Distribution Theory*, Chicago 1971.

E. C. Budd, "Postwar Changes in the Size Distribution of Income in the U.S.," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 247–60.

J. Conlisk, "Can Equalization of Opportunity Reduce Social Mobility?," *Amer. Econ. Rev.*, Mar. 1974, *64*, 80–90.

M. David and R. Miller, "A Naive History of Individual Incomes in Wisconsin, 1947–59," mimeo, Univ. of Wisconsin, June 1968.

Peter B. Doeringer and Michael Piore, *Internal Labor Markets and Manpower Analysis*, Lexington 1971.

Milton Friedman, *Capitalism and Freedom*, Chicago 1962.

George Furstenberg et al., *Patterns of Discrimination*, Lexington 1974.

David Gordon, *Theories of Poverty and Underemployment*, Lexington 1972.

W. J. Haley, "Human Capital: The Choice Between Investment and Income," *Amer. Econ. Rev.*, Dec. 1973, *63*, 929–44.

P. Hart, "The Dynamics of Earnings; 1963–1973," *Econ. J.*, Sept. 1976, *86*, 551–65.

P. Henle, "Exploring the Distribution of Earned Income," *Mon. Lab. Rev.*, Dec. 1972, *95*, 16–27.

R. Herrnstein, "I.Q.," *Atlantic*, Sept. 1971.

Christopher Jencks et al., *Inequality*, New York 1972.

T. Johnson and F. J. Hebein, "Investments in Human Capital and Growth in Personal Income," *Amer. Econ. Rev.*, Sept. 1974, *64*, 604–15.

Andrew I. Kohen, *Career Thresholds*, Vol. 4, Columbus 1973.

E. Lazear, "Age, Experience, and Wage Growth," *Amer. Econ. Rev.*, Sept. 1976, *66*, 548–58.

D. E. Leigh, "The Effect of Job Experience on Earnings Among Middle-Aged Men," *Ind. Relat.*, May 1976, *15*, 130–46.

———, "Occupational Advancement in the Late 1960's: An Indirect Test of the Dual Labor Market Hypothesis," *J. Hum. Resources*, Spring 1976, *11*, 155–71.

F. Levy, "How Big is the American Underclass?," work. paper no. 39, Univ. California-Berkely 1975.

L. A. Lillard, "Inequality: Earnings vs. Human Wealth," *Amer. Econ. Rev.*, Mar. 1977, *67*, 42–53.

——— **and Y. Weiss,** "Analysis of Longitudinal Earnings Data: American Scientists 1960–70," Nat. Bur. Econ. Res. work. paper no. 121, New York 1976.

Harold Lydall, *The Structure of Earnings*, Oxford 1968.

John J. McCall, *Income Mobility, Racial Discrimination, and Economic Growth*, Lexington 1973.

J. Mincer, "The Distribution of Labor Incomes: A Survey," *J. Econ. Lit.*, Mar. 1970, *7*, 1–26.

———, *Schooling, Experience, and Earnings*, New York 1974.

James N. Morgan et al., *Five Thousand Families—Patterns of Economic Progress*, Ann Arbor 1974.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598–609.

Herbert S. Parnes, *The Pre-Retirement Years*, Washington 1970.

Jan Pen, *Income Distribution*, New York 1971.

Michael J. Piore, "Fragments of a Sociological Theory of Wages," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 377–84.

———, *Labor Market Stratification and Wage Determination*, Washington 1974.

S. Rosen, "Learnings and Experience in the Labor Market," *J. Hum. Resources*, Summer 1972, *7*, 326–42.

N. D. Ruggles and R. Ruggles, "The Anatomy of Earnings Behavior," mimeo, Nat. Bur Econ. Res. Conference on Research in Income and Wealth, May 1974.

Bradley R. Schiller, "Class Discrimination vs. Racial Discrimination," *Rev. Econ. Statist.*, Aug. 1971, *53*, 263–69.

———, "Stratified Opportunities: The Essence of the 'Vicious Circle,'" *Amer. J. Soc.*, Nov. 1970, *76*, 426–42.

———, *The Economics of Poverty and Discrimination*, 2d ed., Englewood Cliffs 1976.

T. P. Schultz, "Long-Term Change in Personal Income Distribution: Theoretical Approaches, Evidence and Explanations," RAND, Santa Monica 1972.

J. D. Smith and J. M. Morgan, "Variability of Economic Well-Being and its Determinants," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 286–95.

A. B. Sorenson, "Growth in Occupational Achievement: Social Mobility or Investment in Human Capital," in Kenneth Land and Seymour Spilerman, eds., *Social Indicator Models*, New York 1975.

Lee Soltow, *Six Papers on the Size Distribution of Wealth and Income*, New York 1969.

P. Taubman, "Schooling, Ability, Non-Pecuniary Rewards, Socioeconomic Background and the Lifetime Distribution of Earnings," Nat. Bur. Econ. Res. work. paper no. 17, New York 1973.

Lester C. Thurow, *Generating Inequality*, New York 1975.

______ **and R. Lucas**, "The American Distribution of Income: A Structural Problem" study for the Joint Economic Committee, U.S.Congress, Washington Mar. 17, 1972.

H. M. Wachtel, and C. Betsey, "Low Wage Workers and the Dual Labor Market: An Empirical Investigation," *Rev. Black Polit. Econ.*, Spring 1975, *5*, 288-301.

M. Wachter, "Primary and Secondary Labor Markets: A Critique of the Dual Approach," *Brookings Papers*, Washington 1974, *3*, 637-93.

D. A. Wise, "Personal Attributes, Job Performance, and Probability of Promotion," *Econometrica*, Nov. 1975, *43*, 913-32.

U.S. Department of Commerce, Bureau of Economic Analysis, "Regional Work Force Characteristics and Migration Data: a Handbook on The Social Security Administration's Continuous Work History Sample," Washington 1977.

How Far Can We Push the "Law of One Price"?

By PETER ISARD*

Students exposed to the pure theory of international trade have been seduced by visions of an imaginary world with few goods, each typically produced by several countries but nevertheless homogeneous. In the assumed absence of transport costs and trade restrictions, perfect commodity arbitrage insures that each good is uniformly priced (in common currency units) throughout the world—the "law of one price" prevails.

In reality the law of one price is flagrantly and systematically violated by empirical data. This paper presents evidence that exchange rate changes substantially alter the relative dollar-equivalent prices of the most narrowly defined domestic and foreign manufactured goods for which prices can readily be matched. Moreover, these relative price effects seem to persist for at least several years and cannot be shrugged off as transitory. In other words, for manufactured goods selected from the most disaggregated commodity lists for which *U.S.* and foreign prices can be matched, the products of different countries exhibit relative price behavior which marks them as differentiated products, rather than near-perfect substitutes.

To clarify discussion it is useful to distinguish two contexts in which the law of one price is valid from a third context in which the law of one price does not hold. First, in a comparison of *U.S.*, European, and Japanese prices of various well-defined steel items (plate, galvanized sheet, cold-rolled sheet, and hot-rolled sheet) c.i.f. for delivery in a common port, Laurence Rosenberg found that relative dollar prices charged by different countries were fairly constant over time and were not significantly affected by exchange rate realignments. The dollar prices of primary commodities are also generally considered to be fairly independent of country of origin.[1] These are cases in which the products of *different* countries are close to identical, or near-perfect substitutes, so that any price disparities would be rapidly eliminated by commodity arbitrage. Second, in the absence of restrictions on commodity arbitrage, a product of any *single* country sold competitively in two different markets (foreign or domestic) would also obey the law of one price in the sense that its dollar-equivalent prices in the two markets could not differ by more than the cost of transportation between these markets.

Many *U.S.* manufactured goods do not have near-perfect substitutes on the lists of products manufactured abroad, however, and in this third context the law of one price is denied as an empirical proposition. Agricultural tilling machinery produced in the United States, for example, is apparently not a close substitute for agricultural tilling machinery produced in Germany. More generally, the most disaggregated groupings of manufactured goods for which both *U.S* and German prices are readily available are dominated by products for which German dollar price indexes diverge over time from *U.S.* dollar price indexes[2] in a manner that is strongly correlated with exchange rate movements. This divergence is evident in comparisons of *U.S.* wholesale transactions prices and German export transactions prices for various 2- and 3-digit sectors of the *WPI* industry breakdown (Section I).

*Division of International Finance, Board of Governors of the Federal Reserve System The opinions expressed herein do not necessarily represent the views of the Federal Reserve System. I am grateful to Lance W. Girton, Peter Hooper, and Jeffrey R. Shafer for helpful discussions.

[1]This may not be the case when sellers of primary commodities have monopoly power and/or enter into long-term marketing agreements with their customers, as do *U.S.* copper producers, for example.

[2]This divergence should come as no surprise to anyone familiar with the work of Irving Kravis and Robert Lipsey.

in comparisons of *U.S.* and German export transactions prices for various 4- and 5-digit *SITC* machinery categories (Section II), and in comparisons of *U.S.* export unit values with unit values of *U.S.* imports from Canada, Germany, and Japan for various 7-digit Schedule A and B commodity groups (Section III).

The denial of the law of one price in this context—at the most disaggregated product level for which price data can be readily matched—provides a strong presumption that it is impossible to assemble available data into aggregate price indexes which can be expected to obey the law of one price (except, perhaps, when product coverage is restricted to primary commodities). Obversely, the notion that aggregate indexes of export or tradeable-goods prices will exhibit purchasing power parity—that is, that relative home currency prices of different countries will stay in line with exchange rates—cannot validly lean on the law of one price for support.

I. Comparative Movements of *U.S.* and German Industrial Prices

The adjustment mechanism alleged to police the law of one price is commodity arbitrage. Under free trade, if products were marketed competitively, commodity arbitrage would prevent disparities between the f.o.b. transactions prices associated with export and domestic sales of the same product—that is, export and wholesale transactions prices would be equal f.o.b. International tests of the law of one price would be insensitive to whether the comparisons were between international wholesale prices, export prices, or a mix of both.

Discriminating monopolies and tariffs, subsidies, or other trade restrictions create disparities between export and wholesale prices. Provided that trade restrictions do not change substantially during the data period, however, international comparisons of any mix of export and wholesale prices can validly test the law of one price by focussing on whether any initial disparities change substantially over time. Evidence that disparities between the common currency prices of different countries are systematically correlated with exchange rates, rather than randomly fluctuating over time, is a strong denial of the law of one price for the products being compared.

The United States, Germany, and Japan publish data on export transactions prices. The coverage of *U.S.* data is restricted mainly to various 4- and 5-digit *SITC* machinery items, collected only once a year (in June) prior to 1974 and four times a year beginning in 1974. German and Japanese data are available monthly for a broad list of items; but for many items Japanese prices are sticky. These considerations have led us to first compare monthly time-series of *U.S.* wholesale prices and German export prices for a variety of industries over the 1968–75 period, and to then compare June data on *U.S.* and German export prices for various machinery categories over the 1970–75 period.[3]

The first set of industry price comparisons is described by Figure 1 and Tables 1 and 2. The exchange rate is measured in dollars per mark (1970 = 100),[4] and relative price indexes are German mark prices multiplied by the exchange rate and divided by *U.S.* dollar prices (and then converted to 1970 = 100). The figure presents strong evidence that relative dollar prices of apparel and paper products have not fluctuated about constant levels during the eight-year data period under examination, but rather have been influenced heavily by exchange rate movements. Interpreted casually, the figure suggests that the relative price of apparel is explained almost entirely by the

[3]The first comparison extends a similar study of the 1968–73 period, which I have described elsewhere. This comparison focuses on 2- and 3-digit sectors of the 8-digit *WPI* industry breakdown, and does *not* consider the most narrowly defined industries for which price comparison is possible. The comparisons of machinery export prices (Section II) and *U.S.* import and export unit values (Section III), however, *are* restricted to the most narrowly defined products for which such prices can readily be compared.

[4]For December 1968–September 1969 the exchange rate is set 4 percent above the actual spot rate to reflect the effective exchange rate for German exports under the 4 percent export tax levied between late November 1968 and the mark revaluation in October 1969.

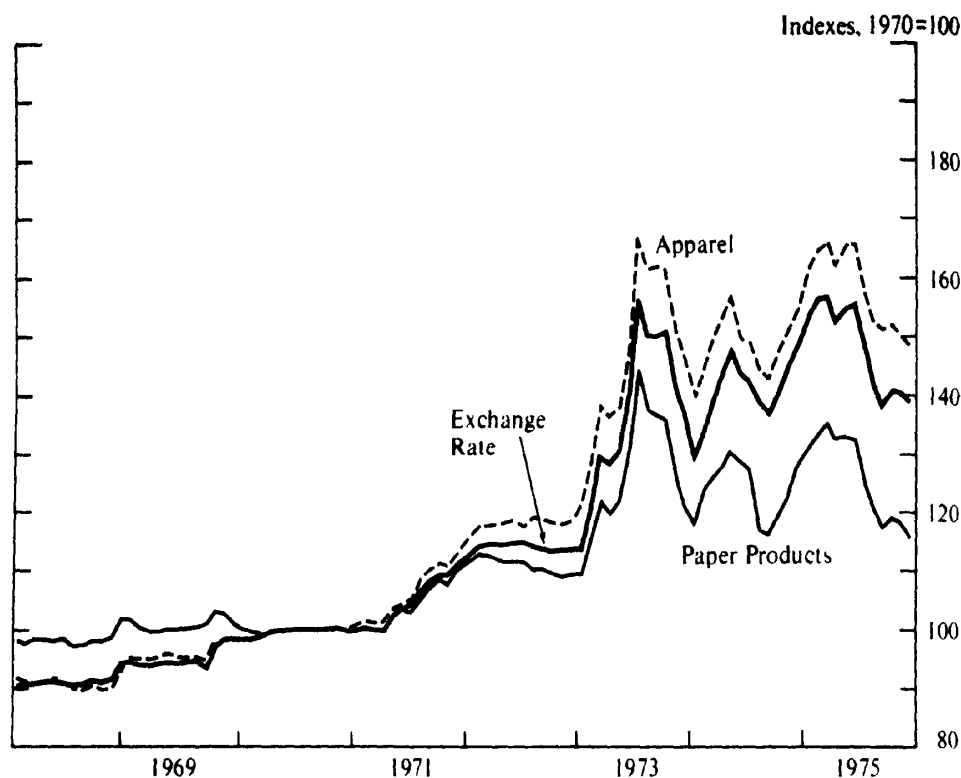



FIGURE 1

TABLE 1—PERCENTAGE CHANGES IN EXCHANGE RATES AND RELATIVE DOLLAR PRICE INDEXES BETWEEN SELECTED PERIODS

	Jan.-Mar. 1968 to June-Aug. 1969	June-Aug. 1969 to Feb.-Apr. 1971	Feb.-Apr. 1971 to July-Sept. 1972	July-Sept. 1972 to Aug.-Oct. 1973	Aug.-Oct. 1973 to Aug.-Oct. 1974	Aug.-Oct. 1974 to Feb.-Apr. 1975	Feb.-Apr. 1975 to Oct.-Dec. 1975	Jan.-Mar. 1968 to Oct.-Dec 1975
Exchange rate (dollars/mark)	4.07	5.85	14.14	31.33	−7.70	12.11	−9.90	53.94
German dollar price/ *U.S.* price								
Apparel	4.57	6.15	16.73	36.26	−10.52	13.61	−8.47	64.29
Industrial chemicals	6.17	3.36	9.91	43.92	−15.01	−13.47	−16.04	7.18
Agricultural chemicals	8.39	−4.51	9.02	37.36	14.37	−10.26	−27.05	16.06
Plastic materials	10.73	10.90	14.93	28.96	−20.23	−10.13	−13.11	13.37
Paper products	2.08	0.86	9.35	23.57	−14.12	13.92	−12.00	19.78
Metalworking machinery	10.68	20.66	14.23	33.52	−14.98	11.08	−12.04	69.18
Electrical industrial equipment	5.36	7.85	15.04	34.08	−5.06	8.88	−11.85	59.71
Home electronic equipment	10.99	8.07	10.87	38.32	−5.68	12.69	−9.22	77.51
Glass products	−0.47	−3.66	19.54	38.97	−11.61	4.44	−13.26	27.55

Source: U.S. price data are from U.S. Department of Labor, *Monthly Labor Review*. German price data are from Statistisches Bundesamt. A detailed data appendix is available from the author upon request.

TABLE 2—CUMULATIVE PERCENTAGE CHANGES IN EXCHANGE RATES AND RELATIVE DOLLAR PRICE INDEXES

	June-Aug. 1969	Feb.-Apr. 1971	July-Sept. 1972	Jan.-Mar. 1968 to Aug.-Oct. 1973	Aug.-Oct 1974	Feb.-Apr. 1975	Oct.-Dec. 1975
Exchange rate (dollars/mark)	4.06	10.16	25.73	65.13	52.41	70.86	53.94
German dollar price/ *U.S.* price							
Apparel	4.57	11.00	29.57	76.55	57.98	79.49	64.29
Industrial chemicals	6.17	9.73	20.61	73.58	47.53	27.66	7.18
Agricultural chemicals	8.39	3.50	12.83	54.99	77.27	59.09	16.06
Plastic materials	10.73	22.79	41.13	82.00	45.18	30.48	13.37
Paper products	2.08	2.96	12.59	39.14	19.49	36.12	19.78
Metalworking machinery	10.68	33.55	52.56	103.69	73.16	92.34	69.18
Electrical industrial equipment	5.36	13.63	30.72	75.26	66.40	81.17	59.71
Home electronic equipment	10.99	19.95	33.00	83.96	73.50	95.53	77.51
Glass products	−0.47	−4.11	14.63	59.30	40.80	47.05	27.55

Source: See Table 1.

xchange rate, whereas the relative price of ıaper products adjusts almost entirely to xchange rate changes in the short run while noving back slowly toward its initial level ıver time.

In Tables 1 and 2 we have selected 8 hree-month periods during which the exhange rate was fairly stable. Table 1 comıares movements in exchange rates and elative prices during the successive interals between these 8 periods, while Table 2 hows cumulative changes. The first interval n Table 1 starts at the beginning of the data ıeriod and ends just prior to the German evaluation in October 1969. The second nterval includes this revaluation and ends ust prior to the start of the German float n May 1971. The third interval spans the Ϟerman float, the Smithsonian Agreement n December 1971, and the three quarters ollowing the Smithsonian. This interval nds prior to the early signs of the pressures hat brought the realignment in February 973. The fourth interval includes the realignments of first-quarter 1973 and the loating period thereafter, ending when the nark was at its peak in summer 1973. The ifth interval ends a year later, after the nark had fallen to a trough in January 1974, risen to a new peak in May, and then depreciated to a summer 1974 trough. The sixth interval ends with the mark at its next peak in spring 1975. The seventh interval spans the dollar appreciation in the second half of 1975.

For most of these intervals, changes in the exchange rate are paralleled fairly closely by movements in five of the nine relative price indexes—those for apparel, metalworking machinery, electrical industrial equipment, home electronic equipment, and glass products—although relative prices of metalworking machinery and home electronic equipment show "unexplained" upward shifts in the first two intervals while the relative price of glass products shows unexplained downward shifts. The relative price of paper products moves up proportionately less (or down proportionately more) than the exchange rate in six out of seven intervals; while relative prices of industrial chemicals, agricultural chemicals, and plastic materials parallel exchange rates fairly closely for the first half of the sample period and then fall sharply during the second half.

The conclusions drawn from this informal analysis are 1) that exchange rate movements are associated with substantial shortrun changes in relative dollar price indexes

TABLE 3—EXCHANGE RATES AND RELATIVE EXPORT PRICE INDEXES FOR SELECTED MACHINERY CATEGORIES[a]

		German Dollar Price/*U.S.* Dollar Price					
	Exchange Rate	Internal Combustion Engines	Agricultural Tilling Machinery	Office Calculating Machines	Metalworking Machinery	Pumps	Forklift Trucks
June 1970	100	100	100	100	100	100	100
June 1971	103.4	104.1	108.9	110.3	110.4	106.2	111.1
June 1972	114.6	119.8	116.6	114.4	125.2	121.2	125.6
June 1973	140.9	155.5	136.2	139.3	153.8	144.7	159.7
June 1974	143.9	147.7	138.1	146.0	144.3	151.7	145.1
June 1975	155.2	148.1	122.5	147.7	141.8	139.3	139.1

Source: U.S. price data are from U.S. Department of Labor, *U.S. Export and Import Price Indexes.* German data are from source listed under Table 1.
[a]1970 = 100.

for all industrial categories considered here, and 2) that in most cases a major share of the short-run relative price change persists for at least several years. Careful econometric studies of data for a longer sample period might indeed find that the relative price changes associated with any particular exchange rate movement are completely offset over long periods of time. But in reality exchange rates are rarely stable over long periods of time. Thus, for practical purposes, products at this level of disaggregation are not sufficiently close substitutes to preclude substantial and persistant changes in relative common currency prices.

II. Comparative Movements of *U.S.* and German Export Prices for Selected Machinery Categories

The *U.S.* and German export transactions prices for various 4- and 5-digit *SITC* machinery categories allow relative price comparisons at a finer level of commodity disaggregation than the industry groups considered above. Prior to 1974, *U.S.* data were collected only once a year, in June. Table 3 compares relative prices for six machinery categories with the exchange rate in June of each year during the 1970-75 period. The conclusions of the previous section extend to this finer level of disaggregation. Machinery items at the 4- and 5-digit *SITC* level of disaggregation are not sufficiently close substitutes to preclude substantial and persistent changes in relative common currency prices.

III. Comparisons of *U.S.* Export Unit Values with Unit Values of *U.S.* Imports from Canada, Germany and Japan

The *U.S.* export price data are available at a still finer level of product disaggregation in the form of unit value indexes for 7-digit Schedule B export commodities. These export unit values can be compared with 7-digit Schedule A import unit values for products distinguished by country of origin.[5,6] Unlike the process of collecting transactions price data, however, the process of collecting unit value information does not hold constant the mix of items within each commodity group whose prices are sampled. Thus, on the one hand, there is no strong presumption that the law of one price will be more evident in these unit value data than it is in the export and

[5]The Schedule A and Schedule B classifications differ, but a reasonably close matching is possible at the 7-digit level.

[6]It is not appropriate to dismiss this comparison on the grounds that countries rarely export the exact same products that they import. The relevant issue is whether products selected from disaggregated lists of *U.S.* manufactured goods have close (not exact) substitutes on lists of goods manufactured abroad, and this is an empirical question which should be addressed in each of the few contexts for which matching data are available.

wholesale price data previously examined for less disaggregated commodity groups. But on the other hand, there is no presumption that shifts in commodity composition will generate "noise" in relative export and import unit values that is strongly correlated with exchange rate movements.

Because our only access to these unit value data was by hand-copying, we limited our sample size to five commodity groups and constructed unit values on a quarterly basis, rather than monthly, from first-quarter 1968 through first-quarter 1975. Our focus is on unit values of exports to all importing areas combined and unit values of imports from three selected countries: Canada, Germany and Japan. The five commodity groups are soaps, tires (pneumatic passenger car), wall paper, ceramic tile (floor and wall), and steel bars. Export unit values are generally f.a.s. at the *U.S.* port of export, based on the transactions price, including inland freight, insurance and other charges incurred in placing the merchandise alongside the carrier at the *U.S.* port of exportation. Import unit values are c.i.f. beginning in 1974; prior to 1974 c.i.f. values are not available and, import value is defined generally as "the market value in the foreign country."

These unit value data fluctuate so erratically that it is difficult to reach any conclusions about the law of one price by looking casually at plots analogous to Figure 1, or at information analogous to that provided in Tables 1–3. Accordingly I have relied on regression analysis to determine if any part of the variation in ratios of import unit values to export unit values is related systematically to fluctuations in exchange rates.

The notation I have used is:

t = index of quarterly time periods

R_t = ratio of *U.S.* import unit value by

TABLE 4—REGRESSION RESULTS FOR HYPOTHESIS (1)

	a_0	a_1	a_2	ρ	$\bar{R}^2$	*D.W.*
Canada						
Tires	−4.16	.0588	−.317	.859	.737	2.33
	(−.832)	(1.16)	(−1.15)	(8.87)		
Wallpaper	−.406	.0118	.361	186	.656	1.89
	(−.462)	(1.31)	(4.50)	(1.00)		
Steel bars	.852	−.00292	.418	.0930	.553	1.94
	(.935)	(−.312)	(4.82)	(.494)		
Germany						
Soap	.726	.0938	−.791	.120	.148	1.60
	(.607)	(2.35)	(−1.35)	(.641)		
Tires	−.0828	.0437	−.142	.758	.728	1.66
	(−.152)	(2 72)	(−.816)	(6.15)		
Wallpaper	.316	.0264	−.0401	−.0974	.163	1.96
	(.885)	(2.21)	(−.223)	(−.518)		
Japan						
Soap	−.582	15.49	.921	.113	.0674	1.87
	(−.137)	(1.12)	(.740)	(.604)		
Tires	−.940	6.28	.244	.461	.869	2.11
	(−2.90)	(6.04)	(2.95)	(2.75)		
Wallpaper	−.720	6.79	1.07	.153	.901	2.04
	(−1.83)	(5.30)	(9.40)	(8.17)		
Ceramic tile	.0242	2.32	.428	.125	.693	1.74
	(.0826)	(2.43)	(4.99)	(.665)		
Steel bars	.183	1.39	.148	.508	.672	2.42
	(.825)	(1.95)	(2.71)	(3.12)		

Source: U.S. Department of Commerce.
Note: Numbers in parentheses are *t*-values. Critical values for the one-tailed *t*-test are 1.71 (95 percent confidence) and 2.48 (99 percent confidence).

country of origin to *U.S.* export unit value, in period t

S_t = exchange rate in period t: the *U.S.* dollar price of one unit of the currency of the country of origin of *U.S.* imports

D_t = dummy variable: 0 from 1968 Q1 to 1973 Q4; 1 from 1974 Q1 to 1975 Q1

where D_t is introduced to adjust for the shift as of first-quarter 1974 in the method of valuing imports.

The first regression hypothesis that we tested is

$$(1) \quad R_t = a_0 + a_1 S_t + a_2 D_t + e_t + \rho e_{t-1}$$

which allows for first-order autocorrelation. We also tested the hypothesis

$$(2) \quad \Delta R_t = b_1 \Delta S_t + b_2 \Delta D_t + u_t + \sigma u_{t-1}$$

which allows for a different pattern of serial correlation.[7] In each case we used the Cochrane-Orcutt procedure. The empirical results argued in favor of hypothesis (1) on two counts: the Durbin-Watson statistics were closer to 2.0 for six out of eleven pairs of commodities and countries of origin,[8] while corrected R^2 statistics were consistently higher.[9]

Table 4 presents the regression results for hypothesis (1). Ratios of German dollar prices to *U.S.* dollar prices and ratios of Japanese dollar prices to *U.S.* dollar prices are seen to be significantly and positively dependent on *U.S.* dollar prices of the mark and yen, respectively, for almost all commodity groups under consideration. A similar finding does not emerge in the Canadian case, perhaps because the exchange rate between the *U.S.* and Canadian dollars showed little variance and no abrupt changes during the sample period. The significance of exchange rate levels in the German and Japanese cases, however, suggests again that substantial changes in exchange rates typically have substantial and persistent effects on the relative common currency prices of closely matched manufactures produced in different countries.

[7]Other patterns of serial correlation are more difficult to take into account and have been implicitly assumed not to exist.

[8]Four of the fifteen pairs were discarded because *U.S.* imports from the country of origin were zero or negligible, so that unit values could not be computed, in one or more quarters of the sample period.

[9]It is worth noting, however, that b_1 was judged to be significantly greater than zero with at least 95 percent confidence in five of the eleven estimates of equation (2).

REFERENCES

P. Isard, "The Price Effects of Exchange Rate Changes," in Peter B. Clark et al., eds., *The Effects of Exchange Rate Adjustments*, Washington 1977.

Irving B. Kravis and Robert E. Lipsey, *Price Competitiveness in World Trade*, Nat. Bur. Econ. Res. *Stud. in Int. Econ. Relations*, Vol. 6, New York 1971.

L. C. Rosenberg, "Impact of the Smithsonian and February 1973 Devaluations on Imports: A Case Study of Steel," in Peter B Clark et al., eds., *The Effects of Exchange Rate Adjustments*, Washington 1977.

Statistisches Bundesamt, *Preise Lohne Wirtschaftsrechnungen: Preise und Preisindices für Aussenhandelsguter*, Wiesbaden, various issues.

U.S. Department of Commerce, *U.S. Exports Schedule B Commodity by Country*, Washington, various issues.

———, *U.S. General Imports: Schedule A Commodity by Country*, Washington, various issues.

U.S. Department of Labor, *Monthly Labor Review*, Washington, various issues.

———, *U.S. Export and Import Price Indexes*, Washington, various issues.

Education and Screening

By KENNETH I. WOLPIN*

Since the advent of the human capital concept, much attention has been devoted to the relationship between income and schooling. Their positive association is one of the most consistent empirical findings of the human capital literature. The conventional view, the productivity augmenting view, is that schooling enhances earnings via the production of marketable skills. But, recent theoretical arguments have suggested the possibility that schooling's private monetary return may be informationally based.[1] In the most extreme form of this screening hypothesis, schooling serves only to identify those individuals who are more productive in the market, the proposition being that an individual's productivity is unaffected by the formal schooling process.

The importance of these competing explanations relates to schooling's implied social return. If schooling's sole function is informational, its social product is determined exclusively by the gain to productive rearrangements which are made feasible by the less imperfect *ex ante* knowledge of individual productivities. In the most extreme form of screening it is only the relationship between aggregate output and schooling's informational content that determines, along with the resource cost of schooling, the socially optimal investment in schooling. It is the value of this information from both a social and private perspective that is the focus of the first part of this paper.

One component of schooling's informational return previously explored in the literature is that of job assignment. If individuals differ in their capabilities according to the jobs they perform, and if substitution of labor is imperfect between jobs, firms would maximize output only by some nonarbitrary assignment of workers to jobs.[2] Aggregate output is therefore larger if workers can be identified and "assigned" to their most productive use. A second component concerns the notion that output at the firm level increases with the homogeneity of the workforce, homogeneity which can only be obtained if workers can be categorized prior to employment.[3] Yet, homogeneity itself is not relevant to the argument, since all that is necessary is for particular combinations of workers to produce more than others. Indeed, there is an obvious similarity between this argument and that of job assignment. A third component, considered in this paper, concerns the direct effect of skill dispersion on factor demand and on the level of aggregate output. If inputs are of uncertain quality then not only may there be an intrafirm misallocation in terms of job assignment, but also an interfirm misallocation in the sense of *ex post* variations in the marginal product of identical inputs. The use of schooling as a screening device reduces both forms of inefficiency, although it is the latter that is emphasized here. More generally, in the following section, a static model of input demand is developed under conditions of imperfect information about the quality of inputs. Within that framework a private

*Department of economics and Institution for Social and Policy Studies, Yale University. This paper is a substantially revised version of part of my doctoral dissertation. I am greatly indebted to James P. Smith, Finis Welch, and Robert J. Willis for their advice and comments. I nevertheless retain sole responsibility for remaining errors. This research was funded, in part, by a dissertation grant from the Manpower Administration of the Department of Labor and through a predoctoral fellowship from the National Bureau of Economic Research. This paper is not an official NBER publication since it has not been reviewed by the Board of Directors.

[1]For a formal presentation of the argument see A Michael Spence (1973) or Joseph Stiglitz.

[2]See Kenneth Arrow, Spence (1974), or Stiglitz.

[3]See Stiglitz. However, he offers no rigorous rationale for the homogeneity argument.

and social rationale for screening is deduced.[4]

The point of the screening interpretation is that the private return to schooling may be generated without relying on the acquisition of market related skills. The related empirical question concerns the extent to which productivity differentials are determined by innate as opposed to schooling-acquired attributes. The difficulty in empirically disentangling the two views is fairly obvious. If productivity could be directly measured prior to employment, there would be no need for screening. Yet, to directly test the screening notion, preschool measures of productivity are necessary. The fact that a researcher has what is believed to be a proxy for innate productivity, or productivity prior to school completion, cannot be used to provide evidence one way or the other unless it is assumed that this or better information is unavailable to firms. Therefore, in the second section of this paper, a different methodological approach is outlined and empirically implemented. It is a test based upon the obvious differential return to the purchase of the screen by groups for whom the purchase is or is not necessary for productivity identification. Individuals who do not have to signal or identify their productivities to anyone else need not spend resources to acquire the signal. As between those of equal innate productivity, individuals from the unscreened group would obtain less schooling than those from the screened group.[5] I approximate the former group by nonprofessional self-employed workers and the latter by nonprofessional salaried workers. The reason for this choice, its associated qualifications, the data source, and empirical results are detailed in a later section.

[4]The private motivation refers to the firm's gain from using the information. Ignored for the most part is the general equilibrium aspect of the problem, i.e., the motivation of individuals to acquire the characteristic upon which firms screen, in particular, the greater incentive for the more productive to purchase more schooling. We merely take as given those assumptions necessary to elicit that outcome (see Spence, 1973, 1974).

[5]John Riley's formalization and extension of this argument concerning screened vs. unscreened groups (jobs) led me to reconsider and correct several points of a previous draft of this paper.

I. The Model

A. *The Demand for Inputs of Uncertain Quality*

The stimulus for job market screening is derived from imperfect information about the quality of prospective employees. Workers must be selected from a population composed of individuals possessing a diverse set of productive attributes most, if not all, of which cannot be observed by firms prior to employment and possibly for some period after. More concretely, let $k_i = (k_{i1}, k_{i2}, \ldots, k_{in})$ be the ith individual's vector of productive attributes or skills where the total potential set consists of n different types, for example, technical knowledge, motivation, manual dexterity, so that for any single individual zero elements are not excluded. Schooling can be viewed as either augmenting the whole set or any particular subset of the skill vector and/or as a predictor of the vector. It therefore does not itself enter as an element of the vector. Other potential information sources such as race, sex, previous market experience, and the like would be treated similarly.

Corresponding to a job task or "occupation" there is assumed to exist a mapping of each individual's combination of attributes into a unique skill index denoted by s_i. Since different jobs may require, in a technological sense, different combinations of skills, and since individuals have different skill vectors, the skill index assigned to any particular individual will differ by occupation. Likewise for any given job the skill index will vary across individuals.

To simplify the analysis we consider only a single occupation together with a homogeneous nonlabor input. The production process within a firm is assumed to take the following form:

$$(1) \qquad Y = F(S, K)$$

where

$S = \sum_{i=1}^{L} s_i$ = aggregate skill (index)

L = the number of individuals employed
K = a nonlabor input of homogeneous quality

where F has the usual neoclassical properties.[6]

To capture the notion of imperfect information, it is assumed that firms have no a priori estimates of any single individual's skill vector. Instead, it is assumed that for the population as a whole the probability distribution of the skill index is known with certainty. Further, the firm is viewed as drawing a random sample from the population, its decision variables being the number of individuals to sample (L) and the amount of the capital input (K) to purchase. Since, with respect to the labor input, the firm is concerned only with the sample skill mean it draws, its relevant frequency distribution can be approximated by a normal variate. Therefore, letting $\bar{s}$ be the obtained sample skill mean, its first two moments are μ and σ^2/L where μ and σ^2 are the known population moments. The firm receives $S = \bar{s}L$ units of the aggregate skill index which is itself distributed with mean $\bar{S} = \mu L$ and variance $V = \sigma^2 L$.[7]

Upon taking a second-order Taylor series approximation to the production function around the point $(\bar{S}, K)$, expected output is given by

$$(2) \quad \bar{Y} = F(\bar{S}, K) + \frac{1}{2}\sigma^2 L \frac{\partial^2 F(\bar{S}, K)}{\partial S^2}$$
$$= F(\bar{S}, K) + \frac{1}{2}\frac{\sigma^2}{\mu}\bar{S}\frac{\partial^2 F(\bar{S}, K)}{\partial S^2}$$

More generally, expected output is characterized by

$$(3) \quad \bar{Y} = \Phi(\bar{S}, K, R)$$

where $R = \sigma^2/\mu$ (the variance-mean skill ratio for the population).[8]

Now, the first effect attributable to the introduction of a heterogeneous workforce is a reduction in expected output at the original equilibrium input levels. An expected profit-maximizing firm would always prefer to sample from a skill distribution characterized by lower variance.[9] It does not follow, however, that such a firm would reduce its employment of labor the more heterogeneous it becomes. There are, in fact, several possibly countervailing effects of increased dispersion on labor demand.

First, at the new lower level of expected output there is a pure substitution effect given by the impact of variance on the marginal rate of substitution between labor and capital. Although one might expect firms to substitute away from an input whose quality uncertainty has increased, this is not necessarily the case. It is possible that increasing the labor input reduces the adverse effect of variance proportionately more than does an equivalent increase in the nonlabor input. The sign of the substitution effect is determined by this comparison. Moreover, this effect depends crucially upon third partial derivatives.[10] The reason is quite simple. Since it is declining marginal productivity that makes skill variance costly, the change in the rapidity with which

[6]For the extension to the multioccupation case, see my dissertation. The generalization assumes a different mapping of the skill vector for each occupation and a production process for the firm given by $Y = F(S_1, S_2, \ldots, S_v, K)$, where there are v occupations and where

$$S_j = \sum_{i=1}^{L_j} s_{ij}$$

is the aggregate level of the skill index from employing L_j workers in the jth occupation.

[7]Note that increasing employment increases aggregate skill variance (V) even though the variance of the sample mean is reduced. There is not necessarily an absolute bias toward drawing large samples though there is one relative to a nonsampling framework.

[8]Note that with normality presumed, equation (3) is a description of any degree polynomial approximation to the expected output function.

[9]This follows from the concavity assumption, since from equation (2), $\Phi_R = (1/2)\bar{S} F_{\bar{S}\bar{S}} < 0$ where $F_{\bar{S}\bar{S}} = \partial^2 F(\bar{S}, K)/\partial S^2$.

[10]The substitution effect is given by

$$\frac{1}{L}\frac{dL}{d\sigma^2} - \frac{1}{K}\frac{dK}{d\sigma^2} = \epsilon_{LK}\left(\frac{\Phi_{L\sigma^2}}{\Phi_L} - \frac{\Phi_{K\sigma^2}}{\Phi_K}\right)$$

where

$$\frac{\Phi_{L\sigma^2}}{\Phi_L} - \frac{\Phi_{K\sigma^2}}{\Phi_K} \propto F_K F_{\bar{S}\bar{S}} + \bar{S} F_K F_{\bar{S}\bar{S}\bar{S}} - \bar{S} F_{\bar{S}} F_{\bar{S}\bar{S}K}$$

and ϵ_{LK} is the elasticity of substitution between L and K, L being evaluated at $\bar{S}/\mu$.

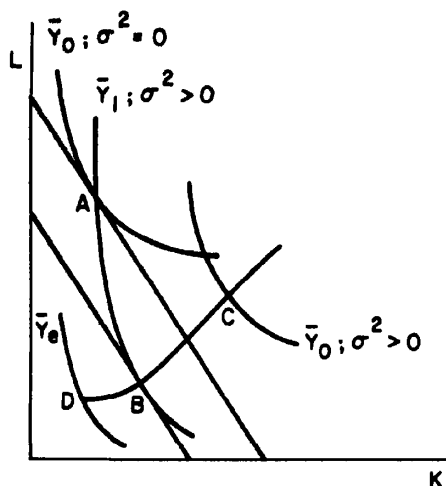



FIGURE 1

it declines as inputs are augmented must be relevant. It is not surprising then that the substitution effect should be more likely to favor the labor input the flatter the marginal product curve becomes, that is, the more positive or less negative is $\partial^3 F/\partial S^3$.

Figure 1 illustrates the optimal input reallocation when skill variance is introduced, in particular the case of a negative substitution effect. A corresponds to the position prior to the introduction of skill variance and B corresponds to the new equilibrium factor ratio established at the lower level of expected output ($\bar{Y}_1$) after introducing variance. There are, however, two further effects. The first is a direct production effect corresponding to a northward movement along the new expansion path in order to restore output to its previous level ($\bar{Y}_0$). This, together with the substitution effect, corresponds to the usual output constant substitution adjustment. Second, there is an induced production effect in response to a change in marginal expected cost after regaining the original output level. In Figure 1, the direct effect is shown as a movement from B to C and the induced effect from C to D. Both are movements along the same expansion path and must be in opposite directions with the net scale effect, that is, the output effect evaluated at B, ambiguous in sign.[11] Although, as depicted in Figure 1, the demand for labor is diminished with the introduction of uncertainty, the demand for labor may rise if the substitution effect favors labor and dominates the scale effect, which will generally reduce the use of all inputs.

B. *The Private and Social Value of Screening*

The role for screening devices is readily apparent from the fact that skill variance reduces expected profits. To illustrate, suppose there is a characteristic of individuals, schooling for example, which segments the total population into subgroups differing in their skill distribution parameters, μ and σ^2. In particular, assume there to be only two groups, E_1 and E_2, with respective means and variances μ_1, σ_1^2, and μ_2, σ_2^2. The *ex ante* information concerning individual schooling attainment enables the expected profit-maximizing firm to sample independently from within each of the two groups. The firm thus makes its employment decision with respect to the number of workers of each group to hire. Its obtained aggregate skill is $S = \bar{s}_1 L_1 + \bar{s}_2 L_2$ with expectation $\bar{S} = \mu_1 L_1 + \mu_2 L_2$ and variance $V = \sigma_1^2 L_1 + \sigma_2^2 L_2$.[12]

From equation (3), expected output becomes

$$(4) \quad \bar{Y} = \Phi\left(\mu_1 L_1 + \mu_2 L_2, \frac{\sigma_1^2 L_1 + \sigma_2^2 L_2}{\mu_1 L_1 + \mu_2 L_2}, K\right)$$

It can be easily demonstrated that the demand for workers from the two groups will depend upon the degree to which their average skill levels, and their variance-mean

[11] The net scale effect is determined by the sign of the following expression

$$\frac{-d\lambda/\lambda}{d\sigma^2} = \left(\frac{\Phi_{L\sigma^2}}{\Phi_L}\frac{E\lambda}{EP_L} + \frac{\Phi_{K\sigma^2}}{\Phi_K}\frac{E\lambda}{EP_K}\right)$$

where λ is marginal cost and $E\lambda/EP_j$ is the elasticity of λ with respect to P_j. See the Appendix for derivations.

[12] Notice that the lack of a covariance term implies that firms do not systematically differ in their sampling capability across schooling groups.

skill ratios diverge. For a fixed labor input ($\bar{L}$), substituting E_1-type for E_2-type workers alters expected output according to

$$(5)\quad \frac{d\bar{Y}}{dL_1}\bigg|_{L=\bar{L}} = (\mu_1 - \mu_2)\Phi_{\bar{S}} + \frac{\mu_1\mu_2\bar{L}}{\bar{S}^2}(R_1 - R_2)\Phi_R$$

where $R_1 = \sigma_1^2/\mu_1$ and $R_2 = \sigma_2^2/\mu_2$. Thus the demand for E_1 workers will increase more relative to that of E_2 workers the greater is their average skill level, the more homogeneous is their group, the greater is the marginal product of labor and the more rapidly it declines.

It would appear that there has merely been a redistribution of income, that is, a change in the relative demand for different workers. However, the use of the screen reduces the effective heterogeneity of the workforce, and must therefore increase the firm's expected output.[13] If skill dispersion reduces labor demand, the use of the screen must in general increase the demand for both groups of labor.[14] It is possible then for both groups to be better off even if schooling is not costless to acquire.[15]

The reduction in effective uncertainty concomitant with the information imparted by schooling necessarily has a positive expected gross social value. Maximization of aggregate output requires the equalization of marginal products across firms.[16] With a perfect screen ($\sigma_1^2, \sigma_2^2 = 0$), the fact that each firm employs the same number of workers from each group is sufficient for aggregate output to be maximized. With an imperfect screen ($\sigma_1^2, \sigma_2^2, \neq 0$), employment of the same number of workers by each firm will not be consistent with the socially optimal distribution between firms since workers are heterogeneous within schooling groups. However, the screen reduces the expected variation in marginal products across firms, and thus will generally increase actual aggregate output. From a social perspective, schooling may have a positive gross social product independent of its productivity augmenting capacity. Whether or not there is too much screening or too little screening will depend not only upon the resource cost of schooling but also upon the cost of variance, Φ_R.[17,18]

[13]The reduction in the variance of the sample mean (V) follows from the assumption of independence in sampling from the two groups. Between-group variance has simply been eliminated, and as long as $\mu_1 \neq \mu_2$, the firm faces less overall uncertainty about the quality of the labor input than it did before the use of the screening device.

[14]Even if the opposite is true, a possibility not ruled out by the previous analysis, the demand for labor might still increase in the long run as new firms enter the industry.

[15]It is an interesting, though tangential, implication of the model that workers of different groups need not be perfect substitutes when $R_1 \neq R_2$. If the relative heterogeneity of the two groups differ, it is possible in equilibrium for the firm to employ some individuals from each group. However, the production function must be a polynomial (or approximated by a polynomial) of greater degree than a quadratic. The reason is simply that expected profit maximization requires that $\Phi_{RR} < 0$, i.e., that the adverse effect of variance on expected output increase with the level of the variance itself. Along with this condition it is necessary that $\Phi_{\bar{S}\bar{S}} < 0$, i.e., that the marginal product of labor decline, and further that $\bar{S}^2\Phi_{\bar{S}\bar{S}}\Phi_{RR} < (\bar{S}\Phi_{\bar{S}R} - \Phi_R)^2$. The proof of these propositions can be obtained from the author as part of a more detailed appendix.

[16]Assume there to be N firms in the industry, each employing L workers. Using equation (2), actual output in the ith firm is

$$Y_i = F(\mu L, K) + \frac{1}{2}L^2F_{\bar{S}\bar{S}}(\bar{s}_i - \mu)^2$$

and aggregate output is

$$\Sigma Y_i = NF(\mu L, K) + \frac{1}{2}L^2F_{\bar{S}\bar{S}}\Sigma(s_i - \mu)^2$$

Since $F_{\bar{S}\bar{S}} < 0$, total product is maximized only where $\Sigma(s_i - \mu)^2 = 0$, i.e., where $\bar{s}_i = \mu$ for all $i = 1, \ldots, N$. Thus, output is maximized where each firm employs the same level of skill, $S = \mu L$, and therefore, where each firm's marginal product of skill is the same.

[17]A general equilibrium model of self-fulfilling prophesies as found in Spence is much more complicated in the context of this model due both to the generality of the production relationship and the sampling framework. I therefore provide no formal proof for the statement in the text that the private return to screening may fall short of the social return. Consider, however, the following intuitive explanation. Suppose that there are only two types of individuals in society with respect to their skill endowments. Further, assume that wages are "set" in such a way that the two groups distinguish themselves by the more productive group purchasing a positive amount of schooling. Schooling is a perfect screen by our definition, since within schooling

II. Empirical Implications and Tests of the Screening Hypothesis

The point of the preceeding section was that any assessment of the social profitability of a screening device must encompass not only its direct contribution to output via the production of human capital, but also its contribution to output via the production of information. Even if the screen is highly productive in the first sense, universal acquisition might not maximize aggregate output if the social cost of identification in terms of resource allocation is sufficiently great. Nor if the information is highly valuable might zero investment in the screen be desirable even if the screen is not causally related to productivity.

As applied to schooling, the basic empirical question is whether the observed schooling distribution of the population corresponds to the "efficient" distribution. To partially answer this question it is necessary to determine the extent to which schooling produces skills as opposed to identifying preexistent skills; a full answer would require knowledge as well of the output cost of input quality uncertainty.

The focus of this section will be on empirical strategies aimed at isolating the innate productivity component of schooling's observed income return. The first approach, the earnings-profile approach, attempts to extract evidence from the schooling-earnings relationship directly. The second approach is a comparison of the schooling decisions of individuals classified into screened and unscreened groups.

A. *The Earnings-Profile Approach*

The fact that screening arises because of imperfect information about productive capabilities prior to hiring implies that earnings should more nearly reflect actual productivities through time as employers learn from direct observation of worker performance. As long as schooling is itself not a perfect screen, the variance in earnings should increase over the life cycle within any given schooling group. Notice, however, that the covariance between earnings and schooling would remain constant if, as in most screening models, schooling accurately predicts the productivity of the average worker. It is also obvious that the simple correlation between earning and schooling would decline while a schooling regression coefficient from a simple regression of schooling on earnings would be constant over the life cycle. All of these results, however, arise solely because of heterogeneity within schooling groups, a

group dispersions are zero. From a social perspective, however, whether or not a perfect screen is warranted depends not only upon the cost of schooling in terms of resource use, but also upon the aggregate output gain from identification. It appears that the social value of schooling could be such that either no screening or full screening is optimal since the latter benefit may or may not be large. If the story is changed slightly so that from the private incentive no screening occurs, say, because the private cost of schooling is very great, the extent to which screening should take place from a social view will depend, as before, also upon the output cost of skill variance It thus seems possible for too little screening to occur. The problem, of course, is that I have not integrated wage formation (and thus the private demand for schooling) into the model and so these conclusions are only tentative. Stiglitz does demonstrate the possibilities for either an over- or underinvestment in the screen, however, in a model which ignores factor demand considerations

[18]If we add the possibility that schooling augments productivities as well, the socially optimal schooling decision becomes more difficult to characterize. However, differentiating the equation for aggregate output given in fn. 16 with respect to the average level of skill in the population yields

$$\frac{d\Sigma Y_i}{d\mu} = NLF_{\bar{s}} + \frac{1}{2}L^3 F_{\bar{s}\bar{s}\bar{s}}\,\Sigma(\bar{s}_i - \mu)^2 + \frac{1}{2}L^2 F_{\bar{s}\bar{s}}\frac{d(\Sigma(\bar{s}_i - \mu)^2)}{d\mu}$$

The first term reflects the direct output effect of the increased aggregate skill level of the population on aggregate output holding screening efficiency constant; the less rapidly the rate of change in the marginal product's decline (the more positive is $F_{\bar{s}\bar{s}\bar{s}}$), the less costly is any imperfection in the screen. The third effect reflects the change in screening efficiency accompanying the rise in average skill. Therefore, the socially optimal investment in schooling depends upon both the direct impact of schooling on skill formation and its indirect impact on the magnitude of schooling's informational content and its value.

fact which is consistent with either the productivity augmenting or screening view.

A somewhat more sophisticated, though similar, approach is proposed by Richard Layard and George Psarcharopoulos. They argue that, with (measured) ability held constant, the schooling regression coefficient from an earnings regression should decline over the life cycle if schooling is merely a screening device. Their theoretical justification relies on employers overassessing the average more-schooled worker, a supposition which is inconsistent with equilibrium screening models. Yet, within the context of the model they propose, the schooling coefficient could be observed to decline over the life cycle without relying on a disequilibrium assumption if the researcher had a better measure of innate productivity than did the employer at the time of initial employment.[19] Indeed, if the researcher had a perfect measure of innate ability, the direct effect of schooling on productivity could be isolated from an earnings regression (containing both schooling and innate ability) performed after firms had become fully cognizant of individual productivities. But it cannot be appropriate to base a test of screening on a presumption that would eliminate the need for the screen, namely a perfect measure of preschool productivity. Further, an imperfect innate ability measure can only be useful, and then only in establishing the existence of screening, if employers are assumed to be less informed than the researcher, for it is only then that the schooling coefficient must decline over the life cycle. Even ignoring these points, the possibility that schooling and ability are correlated with postschool human capital accumulation would confound any interpretation of those regression coefficients, and thus reduce the applicability of this approach.[20]

B. *A Comparison of Screened and Unscreened Groups*

The major implication of the screening model is that more schooling is privately purchased than would be the case if firms could costlessly determine productivities prior to hiring. It therefore follows that if some individuals are employed in jobs in which it is possible to determine productivity at small cost, those individuals should purchase less schooling than their equally skilled counterparts. As long as preschool productivity is a large component of postschool productivity, individuals who, for whatever reason, need not identify their productivities before employment have less incentive to acquire schooling. Therefore, for given innate productivity, the unscreened worker will acquire less schooling than the screened worker or, conversely, for given schooling, the unscreened worker will be of greater innate productivity, and thus have greater earnings than the screened worker.[21]

The problem in applying this test is that of identifying an unscreened group. One possible group consists of the self-employed since they at least need not demonstrate their capabilities to prospective employers. Of course, there may be others, for example, customers, who would provide the incentive for identification. Moreover, the self-employment decision is not necessarily made prior to school completion and those who ultimately become self-employed may hedge by purchasing schooling as insurance in the event they enter the salaried sector. A crude attempt is made to deal with these

[19]What is necessary is that the covariance between the researcher's measure of innate productivity and the firm's assessment of actual productivity (reflected in earnings) increase through time as the firm observes worker performance.

[20]In the empirical work of Paul Taubman and Terence Wales, cited by Layard and Psacharopoulos, both the coefficient on schooling and measured innate ability are seen to rise over the life cycle, a finding that can be interpreted as an on-the-job training effect. Unless the magnitude of this latter component is known by the researcher, running an earnings regression even after full employee identification is acheived could lead to an erroneous conclusion concerning the validity of the screening hypothesis.

[21]These propositions are formally demonstrated by Riley within a signalling framework. The model assumes that the choice between the screened and unscreened job is endogenous, and thus that net discounted lifetime earnings are equalized for the two groups.

issues by considering only a subsample of the self-employed and salaried populations. Since we need to compare a group of self-employed workers whose product is of easily assessed quality to a group of salaried workers whose productivity is difficult to observe, we have limited attention to non-professional workers; the productivity of a self-employed doctor, for example, is no easier to assess than that of a salaried doctor. With respect to the rest of the population, the basic assumption is that the quality of the products or services sold by the self-employed are less costly to determine than the quality of the labor input hired by firms. To partially correct for hedging, individuals were selected from those who were either self-employed or salaried throughout their observable working life. In particular, using the NBER-Thorndike sample, only those whose first and last reported jobs (spanning over twenty years) were as self-employed were compared to a similarly stable class of salaried workers.[22] The argument is simply that those self-employed individuals were more certain of their future employment path, and therefore less likely to purchase schooling for motivations other than human capital investment.

Descriptive statistics for the self-employed and salaried workers satisfying the previous criteria are given in Table 1. The difference in average schooling is .6 years which, given that all individuals in the sample are at least high school completers, implies that the self-employed acquired roughly 75 percent of the amount of extra schooling acquired by the salaried workers.[23] If, as suggested by the ability measure, the two groups are of similar preschool productivity with the self-employed slightly less capable on average, and if the self-employed vs. salaried comparison approximates an unscreened vs. screened comparison, the results suggest that schooling has only a minor screening function. If, instead of using the ability measure as an index of innate productivity, the earnings data are used for that purpose, the screening function takes on greater relevance. This follows from the fact that the two groups obtain about the same schooling even though the self-employed are more able as evidenced by their higher lifetime earnings. Since self-employment earnings most surely include a nonlabor component, the difference in labor earnings is exaggerated and at least the direction of the bias in this comparison is clear.

The validity of the actual test described above is clearly not free of subjective judgements. The choice of the self-employed as an unscreened group relative to salaried workers is not obvious and possibly other dichotomies would be more valuable to explore. However, this kind of comparison appears to be a more fruitful approach than any of those previously suggested.

TABLE 1—MEANS AND STANDARD DEVIATIONS OF SELECTED VARIABLE FOR SALARIED AND SELF-EMPLOYED WORKERS[a]

	Salaried		Self-Employed	
	Mean	Standard Deviation	Mean	Standard Deviation
Schooling	14.55	1.92	13.95	1.81
Ability[b]	−.190	1.72	−.234	1.68
Earnings	8869	6157	12355	10507
Experience	10.98	8.52	11.23	9.31

[a]These figures are based on 3920 life cycle points for salaried workers and 463 for self-employed workers. The life cycle points are formed from observations on 1099 and 157 individuals, respectively. Whether individuals or life cycle points are used as the observational unit is practically irrelevant in the case of schooling and ability, the two variables which have no life cycle variation

[b]The ability measure is a composite of seventeen separate tests conducted in 1943 and is supposed to approximate an IQ-type test.

[22]The NBER-Thorndike sample contains approximately 5000 men who were air force pilot, navigator, and bombardier candidates in 1943. Each individual reported a complete job history between 1945 and 1970. A more detailed description of the data may be found in the work of Taubman and Wales.

[23]A frequency distribution tells much the same story. Approximately 30 percent of the self-employed stopped after high school graduation, an additional 42 percent had some college, and 24 percent graduated from college. The figures for salaried individuals are 23, 35, and 33 percent, respectively.

III. Conclusions

The possibility that schooling performs some identification function with respect to initial capabilities is as difficult to deny as the proposition that schooling enhances those innate capabilities. The real issue concerns not the mere existence of one or the other effect, but the extent to which schooling performs each of these roles. Equally important, however, is a determination of the social product of the information which schooling imparts due to its sorting function. For, as I have attempted to demonstrate in the preceeding sections, the determination of optimal schooling levels viewed from a social perspective must incorporate both of these sometimes conflicting considerations.[24] It is not clear that individuals necessarily overinvest in schooling if it is only a screen, although the likelihood of this occurrence increases as preschool and postschool productivities converge. Thus, one would like to know the proportion of an individual's productivity that is not schooling induced.

The problem in performing the appropriate calculation is that productivity is not easily measured. Even after firms have fully recognized individual differences so that earnings accurately reflect productivity, the researcher still has only an imperfect measure of preschool productivity. Conclusions reached from approaches which make use of ability-based measures of innate productivity are inherently suspect. For this reason, a somewhat different methodology was adopted, namely a comparison of screened and unscreened groups. With some important qualifications, the fact that self-employed workers in nonprofessional occupations obtained about the same level of schooling as nonprofessional salaried workers was taken as evidence against a predominant screening interpretation.

[24]Note that in the present model, the value of the information is itself a function of the degree to which schooling increases skills.

Appendix

Consider the production function given by equation (1) in the text. First-order conditions for profit maximization are

$$\text{(A1)}\quad \overline{Y} = \Phi(\overline{S}, R, K),\ \overline{S} = \mu L,\ R = \sigma^2/\mu$$

$$\text{(A2)}\quad P_L = \lambda\Phi_L$$

$$\text{(A3)}\quad P_K = \lambda\Phi_K$$

$$\text{(A4)}\quad \lambda = P_Y = \text{marginal cost}$$

where P_L and P_K are per unit input prices of labor and capital, respectively. Totally differentiating (A1) to (A3) with respect to σ^2 and rewriting in matrix form

$$\text{(A5)}\quad \begin{pmatrix} 0 & \Phi_L & \Phi_K \\ \Phi_L & \Phi_{LL} & \Phi_{LK} \\ \Phi_K & \Phi_{KL} & \Phi_{KK} \end{pmatrix} \begin{pmatrix} d\lambda/\lambda/d\sigma^2 \\ dL/d\sigma^2 \\ dK/d\sigma^2 \end{pmatrix} = \begin{pmatrix} d\overline{Y}/d\sigma^2 - \Phi_{\sigma^2} \\ -\Phi_{L\sigma^2} \\ -\Phi_{K\sigma^2} \end{pmatrix}$$

The substitution effects, holding output constant at its lower level due to the increased variance, i.e., setting $d\overline{Y}/d\sigma^2 - \Phi_{\sigma^2} = 0$, are

$$\text{(A6)}\quad \frac{1}{L}\frac{dL}{d\sigma^2} = \alpha_K\epsilon_{LK}\left[\frac{\Phi_{L\sigma^2}}{\Phi_L} - \frac{\Phi_{K\sigma^2}}{\Phi_K}\right]$$

$$\text{(A7)}\quad \frac{1}{K}\frac{dK}{d\sigma^2} = -\alpha_L\epsilon_{LK}\left[\frac{\Phi_{L\sigma^2}}{\Phi_L} - \frac{\Phi_{K\sigma^2}}{\Phi_K}\right]$$

where α_j is the cost share of the jth input and ϵ_{LK} is the elasticity of substitution between L and K. The net scale effect is found by solving for $(d\lambda/\lambda)/d\sigma^2$ with $d\overline{Y}/d\sigma^2 - \Phi_{\sigma^2} = 0$.

With segmentation of the population into, say, two schooling groups with skill distribution parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), expected output becomes

$$\text{(A8)}\quad \overline{Y} = \Phi\left(\mu_1 L_1 + \mu_2 L_2, \frac{\sigma_1^2 L_1 + \sigma_2^2 L_2}{\mu_1 L_1 + \mu_2 L_2}, K\right)$$

where it is assumed that firms sample inde-

pendently within subpopulations. With factor employment fixed, $L_1 + L_2 = \bar{L}$ and $K = \bar{K}$, the demand for group one workers is reflected in

$$\text{(A9)} \quad \frac{d\bar{Y}}{dL_1} = (\mu_1 - \mu_2)\Phi_{\bar{S}} + \frac{\partial R}{\partial L_1}\Phi_R = (\mu_1 - \mu_2)\Phi_{\bar{S}} + \frac{\mu_1\mu_2\bar{L}}{\bar{S}^2} \cdot (R_1 - R_2)\Phi_R$$

where $R_1 = \sigma_1^2/\mu_1$ and $R_2 = \sigma_2^2/\mu_2$. There will, thus, be a greater demand for workers from more skilled and more homogeneous groups.

REFERENCES

K. J. Arrow, "Higher Education as a Filter," *J. Publ. Econ.*, July 1973, *2*, 193–216.

R. Layard and G. Psacharopoulos, "The Screening Hypothesis and the Returns to Education," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 985–98.

J. G. Riley, "Information, Screening and Human Capital," work. paper no. 64, Univ. California-Los Angeles 1975.

A. M. Spence, "Job Market Signalling," *Quart. J. Econ.*, Aug. 1973, *87*, 355–79.

———, "Competitive Optimal Responses to Signals: An Analysis of Efficiency and Distribution," *J. Econ. Theory*, Mar. 1974, *7*, 296–332.

J. E. Stiglitz, "The Theory of 'Screening,' Education, and the Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

P. J. Taubman and W. J. Wales, "Higher Education, Mental Ability, and Screening," *J. Polit. Econ.*, Jan./Feb. 1973, *81*, 28–55.

K. I. Wolpin, "Education, Screening and the Demand for Labor of Uncertain Quality," unpublished doctoral dissertation, City Univ. New York 1974.

Trade as Aid: The Political Economy of Tariff Preferences for Developing Countries

By RACHEL MCCULLOCH AND JOSE PINERA*

The use of generalized and nonreciprocal tariff preferences favoring the manufactured exports of developing countries has gained wide support since the proposal was put before the first United Nations Conference on Trade and Development in 1964. Most major developed countries had instituted systems of preferential access by the early 1970's. The United States, initially opposed to the preference concept, finally passed legislation permitting establishment of its own preference scheme as part of the comprehensive Trade Reform Act of 1974. Despite important restrictions on product coverage, tariff preferences now constitute a significant feature of trading relations between developed and less developed nations.[1]

This paper evaluates the use of tariff preferences to generate international resource transfers. The analysis assumes the initial developed country tariff is "rational"—that it is designed to achieve a particular domestic policy objective. Motives for extending preferences are explored in terms of an interdependent developed country welfare function which includes as an argument the flow of resources to the developing world. The policy role of the initial tariff then helps to determine the cost of generating resource flows via preferences in comparison to other assistance channels.

The effects of discriminatory tariff reduction have been discussed extensively in the context of customs union. The traditional analysis emphasizes static efficiency gains and losses, which correspond to the now familiar concepts of "trade creation" and "trade diversion" first used by Jacob Viner.[2] More recently, general equilibrium techniques have been applied to the analysis of customs union.[3] However, neither approach seeks to relate the effects of customs union to the motive for the initial level of protection.[4] The analysis thus fails to explain why a country is initially away from free trade, or why it stops short of unilateral elimination of all tariffs.[5] This criticism applies equally to the analysis of tariff preferences; a logically consistent treatment must take into account the motive for protection. In this paper we assume the initial tariff is rational in the sense that prior to some parameter shift the country enjoys maximum welfare as indicated by the interdependent social welfare function described below. The parameter shift—a change in

*Harvard University and Catholic University of Chile, respectively. This research was supported in part by a grant from the National Science Foundation.

[1] Basic features of the major national schemes are outlined in the *IMF Survey*. On the significance of restrictions limiting preferential access, see Richard Cooper, McCulloch, and Tracy Murray.

[2] Key contributions include those of Viner, James Meade, Harry Johnson (1962), and Richard Lipsey See Melvin Krauss for other references.

[3] Lipsey extended Meade's analysis to take account of large changes in tariff rates. Lipsey explicitly considered three goods, thus allowing for relations of complementarity in consumption Jaroslav Vanek and Murray Kemp have applied modern techniques of general equilibrium analysis to preferential trading. Also see Richard Caves

[4] C A Cooper and B F. Massell (1965a,b) appear to have been the first to raise this objection in the context of customs union, although the use of tariffs to achieve noneconomic objectives was analyzed earlier by Johnson (1960)

[5] In analyzing the effects of preferential tariff cuts, most writers recognize possible gains through improved terms of trade with third countries. However, the question of whether the initial position represents an optimum nondiscriminatory tariff (i.e., one which extracts maximum gains from improved terms of trade for a single tariff on imports from all sources) is not usually raised. It should be noted that further terms of trade gains can accrue through tariff discrimination even when the initial tariff is "optimum" among nondiscriminatory tariffs. This point is elaborated in Section III below.

the welfare function itself or in the possibilities for tradeoffs among its arguments—is seen as the stimulus for extension of tariff preferences.

Since preferential access to developed country markets has been advocated as one of a number of measures designed to assist the development of poor nations, preferences should be compared with other means of achieving development goals. In terms of these goals, preferences favoring exports of manufactured goods have two principal effects on developing economies. First, the resulting change in relative profitability induces *reallocation* of factors of production from traditional industries to manufacturing. Under plausible assumptions this sectoral reallocation facilitates development.[6] However, quite apart from any longer term gain realized through reallocation of factors between industries, preferences assist developing nations by making additional resources available to them. These additional resources, the increase in foreign exchange earnings less the cost of increased exports, can be viewed as a *transfer* from developed to developing countries, and thus comparable in some respects to aid in the form of grants or loans. This paper focuses on the role of preferences in generating such transfers of resources to developing countries, the level of resource transfer—"aid" component—implied by a system of preferences, and the relative cost of transferring resources via preferences rather than through other types of aid.

In Section I of the paper, motives for the initial tariff and for extension of preferences are discussed in terms of the interdependent developed country social welfare function. Section II describes the effects of preferences when the initial tariff provides desired protection for a domestic import-competing industry. The case in which the initial tariff is "optimum" is presented in Section III.

I. Motives for Protection and the Interdependent Social Welfare Function

To formalize the assumption that the initial tariff is consistent with rational government behavior, we assume the developed country maximizes an interdependent social welfare function W with at least three arguments:

$$W = W(I, N, T)$$

The first argument I is real national income, which can be measured by the usual utility function approach; this is often the only argument in the implicit social welfare function used by trade theorists. The second set of determinants of social welfare considered here and represented by N are "noneconomic" objectives such as the level of output, profits, or employment in a particular industry. Income distribution is likely to be among the noneconomic considerations if no cost free internal transfer mechanism exists.[7] The third argument T measures the resource flow to the developing world, which can be viewed as a proxy for the resulting increase in real income or social welfare. The presence of T in the developed

[6] Preferential tariff elimination is similar in its effects to a nondiscriminatory tariff accompanied by a subsidy to developing country manufactured exports, or, equivalently, a subsidy to developing country production and a tax on developing country consumption of manufactures, both at the most favored nation (MFN) tariff rate The conditions under which an export or production subsidy for manufactures will result in a welfare gain for the developing countries and for the world as a whole have been widely discussed. See Johnson (1965).

In the case of present preference arrangements, the importance of "dynamic" considerations is reduced by the limited time for which preferences are to be extended ten years in most cases and the uncertainty surrounding continued preferential treatment for any particular good or country within that period For example, the *U.S.* scheme allows preferences to be discontinued without prior notice for a number of reasons, including the usual escape clause considerations (i.e., developing countries responding too enthusiastically to opportunities generated by preferences). Also, eligibility of goods produced in part from imported inputs depends on the value-added requirement, another parameter subject to change during the life of the system. See Johnson (1966) for further discussion of preferences as a transfer channel

[7] Because both income and its distribution enter into W, an increase in national income as conventionally measured (I) need not lead to an increase in W if the distributive consequences of the increment are associated with a fall in N. In particular, the positive impact on I of a tariff reduction could be offset by negative effects on N.

country welfare function provides the necessary justification for policies which result in resource flows to developing countries at the expense of reductions in I or N. Possible motives for the inclusion of aid flows in W range from genuine altruism—satisfaction derived from the well-being of others—to the national defense considerations which appear to underlie much *U.S.* aid. There may also be the expectation of some implicit compensation, such as more favorable treatment from raw material producers or a better climate for foreign private investment. Other variables could also be included in W; our approach is intended as illustrative and does not attempt to treat comprehensively all factors which may bear upon commercial policy decisions.

Of course, no such explicit function is actually used in the formation of policy; adoption of new policies and retention or discontinuation of old policies reflect a far more complex reality. Nevertheless, policymakers must continually weigh national income effects against compositional or distributive consequences or foreign benefits against domestic costs to reach an overall judgment as to whether a particular change constitutes a welfare improvement in some relevant sense. While the simple welfare function used here obviously cannot capture the full dynamics of the political process, it nonetheless provides a useful framework for analyzing tradeoffs between conflicting national objectives.

The initial tariff is assumed to contribute to developed country welfare through its effects on I, N, and T. Below we consider two specific possibilities: that the tariff protects a desired level of production in a particular industry (Section II); and that the tariff is an optimum tariff which increases real national income I through its effects on the terms of trade (Section III). Assuming that the initial values of the arguments yield a maximum value for W, the change of policy expressed in the decision to institute preferences can be rationalized in two possible ways. The social welfare function may itself have changed. The function W formalizes a process of weighing aid flows against real income and the achievement of domestic goals; as the weights change, so do the social welfare-maximizing values of the arguments. Changed weights may reflect such diverse factors as increased public awareness of foreign poverty, heightened fear of political instability, and threats of raw material cartelization. A higher marginal contribution of aid to social welfare will result in new aid flows at the expense of one or both of the other arguments in W.

A second reason for a policy shift may be seen in terms of the available "technology" for transformations among I, N, and T; that is, changes in the marginal real income cost of increasing N or T. For example, waivers recently obtained now make preferences favoring developing country exports acceptable within the framework of the General Agreement on Tariffs and Trade (*GATT*). Previously such discriminatory arrangements were prohibited unless protected by a grandfather clause. Assuming that countries abide by *GATT* rules, the waiver changes the cost of providing a dollar of aid through preferences. Preferences may also be instituted in place of other aid channels, with total aid constant or decreasing, if *relative* costs associated with these channels are changing over time. The present importance of preferences reflects both rising political costs for donor countries of direct grant and loan aid along with an increased desire on the part of developing nations to receive aid in the form of multilateral untied flows.

II. Preferences and Protection of Domestic Industry

Preferential trade generally makes additional resources available to developing countries.[8] When the additional resources entail a tradeoff between I or N and T for the preference-granting country, these re-

[8]This will be true as long as the supply of exports from developing countries is not infinitely elastic. On the relationship between benefits generated and developing country supply conditions, see Richard Blackhurst (1971).

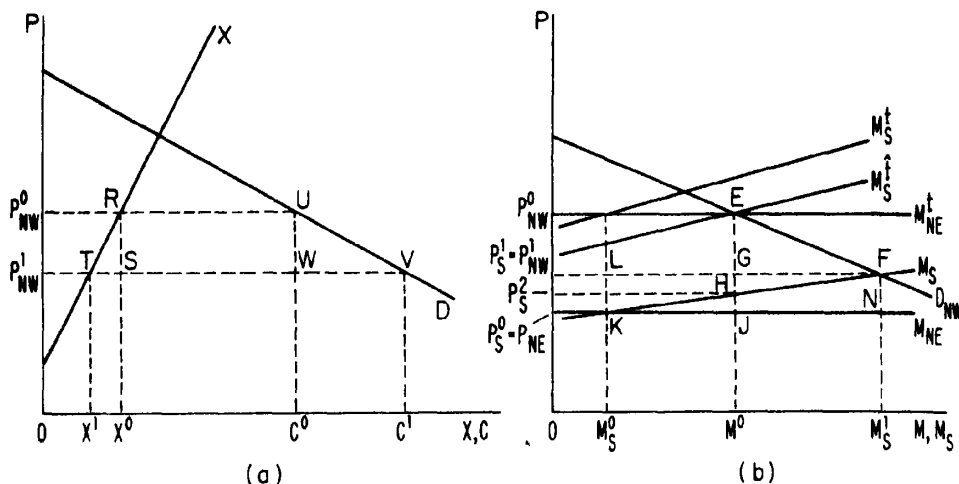



FIGURE 1

sources are here termed aid.[9] The resource value of preferences to a beneficiary country is the increase in export earnings less the social cost of increased exports. Preference-generated aid in this sense will be positive even when developing country supply elasticities are low or zero, as long as exporters receive a higher net price as a result of the preferences.[10]

Figure 1 illustrates the effects of tariff preferences in a world of three countries (or trading blocs). A nondiscriminatory tariff is initially used by Northwest (NW) to maintain a desired level of import-competing production.[11] NW now grants preferences to South (S) while maintaining the initial tariff rate on its trade with Northeast (NE).[12] For simplicity, it is assumed that the price of imports from NE does not depend on the level of NW's purchases.[13] This would be true if production in NE were subject to constant costs, or if NW's imports were small relative to total NE production. Markets are assumed to be competitive.

In Figure 1b, D_{NW} shows NW's net demand for imports from both sources. For each price P_{NW},

$$D_{NW}(P_{NW}) = D(P_{NW}) - X(P_{NW})$$

where D and X are domestic demand and supply in NW, shown in Figure 1a. The schedules M_{NE} and M^t_{NE} show the free trade and tariff inclusive supply to NW of imports from NE; the tariff rate on imports from NE is $(P^0_{NW} - P_{NE})/P_{NE}$. The level of NW's demand for imports from S is then given by $P^0_{NW}ED_{NW}$, which indicates that S can sell up to M^0 at a tariff inclusive price of P^0_{NW}, and larger amounts at correspondingly lower prices. The schedules M_S and M^t_S show the free trade and

[9] This definition implies that any move toward free trade results in aid. Since the original tariff places the protecting country in a position preferred to free trade, movements away from that position motivated by their presumed effect upon welfare in the developing countries are appropriately viewed as aid. However, the original tariff may not be superior to discriminatory trade, even in terms of I alone. Thus, there need not be a tradeoff between I or N and T. In this case, both the preference-granting and the preference-receiving countries can gain, so that the resource transfer to the latter does not constitute aid. This possibility is analyzed in Section III.

[10] This condition will be satisfied if export markets are competitive. However, some observers fear that powerful multinational traders will be able to capture the entire margin between developing country cost and developed market price.

[11] Use of a tariff is taken to indicate that it is preferred by policymakers to a direct subsidy for this purpose. For a discussion of disbursement costs and other considerations which may promote the choice of a tariff over a direct subsidy, see W. M. Corden.

[12] This model draws on Johnson (1962), Blackhurst (1972), and Cooper and Massell (1965b).

[13] This assumption is relaxed in Section III.

tariff inclusive supply of exports to NW from S. Prior to extension of preferences, NW imports a total of $M^0 = C^0 - X^0$, of which M_S^0 is provided by S.

After preferences are instituted by NW, imports from S are allowed tariff free entry, while the original tariff is maintained on imports from NE. As a result, NW's imports from S increase to M_S^1, completely displacing NW's imports from NE.[14] Preferences reduce the domestic price to P_{NW}^1, increase consumption by $M_C = C^1 - C^0$, and decrease import-competing production by $M_X = X^0 - X^1$. Aid to S is the difference between the increase in export receipts and the increase in export costs:

$$P_S^1 M_S^1 - P_S^0 M_S^0 - \int_{M_S^0}^{M_S^1} P_S dM_S$$

If demand and supply curves are linear, this becomes

$$\Delta P_S M_S^0 + 1/2 \Delta P_S (M_X + M_C + M_{NE}^0)$$

corresponding to the area $P_{NW}^1 FKP_{NE}$ in Figure 1b. The net reduction in NW's real income is

$$\Delta P_S M_S^0 + \Delta P_S M_{NE}^0 + 1/2 \Delta P_{NW} (M_X + M_C) \qquad (\Delta P_{NW} < 0)$$

which corresponds to the area

$$P_{NW}^1 LKP_{NE} + LGJK - EFG$$

in Figure 1b.[15] Because of the assumed infinitely elastic supply of imports from NE, NE's welfare is unaffected.[16]

The gain to S less the real income cost to NW is

$$-1/2 \Delta P_{NW}(M_X + M_C) + \Delta P_S(M_X + M_C) - 1/2 \Delta P_S(M_X + M_C + M_{NE}^0)$$

which corresponds to the area

$$EFG + GFNJ - KFN = EFH - HJK$$

in Figure 1b. EFH measures the efficiency gain to NW from expansion of the total imports plus the producer's surplus in S associated with production of those additional imports, while HJK measures the increased real cost of production incurred when imports from S are substituted for lower cost units originally supplied by NE. In the case illustrated by Figure 1, the net impact of these effects is positive; the gain to S outweighs the real income foregone by NW. The analysis thus implies that it may, in effect, cost NW less than \$1 to grant \$1 in aid to S via preferences[17] when preferences represent the only available means of reducing tariffs. In this instance, preferences can be more efficient than direct grants or loans as a means of transferring resources to developing countries.

However, the discussion has not yet taken into account the motive for the initial level of protection. Because the original tariff policy achieved some domestic goal—here, the maintenance of a desired level of output in the import-competing industry—the analysis *overstates* the gain from preferences by ignoring the reduction in NW welfare resulting from lower production in that industry. In terms of the function W, the analysis has looked only at effects on I and T, omitting from consideration the induced change in N. To emphasize the importance of the omission, we reformulate the analysis taking X^0, the original level of import-competing production in NW as *given*,[18] and consider the results of institut-

[14]Complete elimination of exports from NE follows from the assumption of an infinitely elastic supply of imports from NE at a tariff inclusive price above P_{NW}^1.

[15]The loss of tariff revenue is not considered in this calculation. If use of substitute revenue sources results in further reductions in I, these should also be included here.

[16]Costs of transitional adjustment are not taken into account. See articles by Stephen Magee and Malcolm Bale for estimates of trade adjustment costs for the United States.

[17]This measures the *average* cost of the total transfer, using a preferential tariff of zero However, as Lipsey and others have noted, a partial preference is more likely to increase world income. in terms of the discussion here, the ratio of marginal aid to marginal cost incurred by NW falls as the degree of preference is increased. However, the lower is the preferential margin, the smaller the resource transfer to S. This suggests that to produce a given level of resource transfer, partial tariff preferences for many products are superior to complete exemption from tariffs for a few products. See Pinera.

[18]Such a simplification would be consistent with a W function which places great weight on incremental import-competing production up to the level X^0, with

ing tariff preferences in such a way as to continue the same level of protection for the domestic industry. The requirement that X remain fixed implies that P_{NW}, C, and total imports also remain unchanged. This in turn means that only trade diversion can result from provision of tariff preferences.

Depending on supply and demand conditions and the height of the initial tariff, duty free entry into NW of imports from S may allow preferred imports to exactly displace imports from NE, to replace only a fraction of NE imports, or to exceed the previous level of total imports (as in Figure 1b). In the first case, $M_X + M_C = 0$; P_{NW}, X, and C are unaffected. Then $\Delta P_S = P^0_{NW} - P^0_S$, so that resulting aid to S is

$$\Delta T = \Delta P_S M^0_S + 1/2 \Delta P_S M^0_{NE}$$

and the cost in foregone real income to NW is

$$\Delta I = \Delta P_S M^0_S + \Delta P_S M^0_{NE}$$

Thus, conventionally measured world income falls by $1/2\ P_S M^0_{NE}$. Whether NW's social welfare rises or falls as a result of the preference scheme depends on the specific form of the W function. If W rises, the preference scheme increases world welfare despite the decline in world real income as usually measured, a result which reflects a high (marginal) cost of grant aid. For the second case, incomplete trade diversion, the analysis is similar. Again, world income as measured must fall.

The most interesting case is that in which duty free imports from S would exceed the previous level of total imports, as illustrated in Figure 1b. This is inconsistent with the assumed domestic objective of maintaining production at X^0. However, NW could continue protection of the domestic industry at the previous level by allowing S a partial preference, so that the tariff inclusive supply curve M^t_S passes through point E. Calculation of aid to S and cost to NW is now analogous to the first case. Aid to S (ΔT) is measured by $P^2_S HKP^0_S$. The net loss in world real income is HJK. A more comprehensive measure of world welfare may rise or fall depending on the relative weights of T and I in the W function.

As an alternative to a partial preference for imports from S, NW could combine a preferential tariff rate of zero with a quota limiting preferred imports to M^0, the initial level of total imports. If import licenses are provided free to S exporters, or if "voluntary" export quotas are administered by S, aid to S would increase by $P^0_{NW}EHP^2_S$ relative to the partial preference alternative with an equal rise in the cost to NW. This tariff-quota scheme is equivalent to an arrangement in which NW gives to S the receipts from the preferential (but nonzero) tariff on imports from S considered above. It is noteworthy that a system of tariff quotas in which additional imports enter at the original tariff rate will generate more aid than a smaller tariff cut with unrestricted imports, when both are designed to maintain the original level of total imports. The European Community and Japan use tariff quotas to limit preferred imports; however, because licenses are awarded to domestic importers, the system is unlikely to confer extra benefits on developing country exporters.

III. Preferences and Optimal Price Discrimination

A country which has some monopsony power in world markets—i.e., faces an upward sloping supply curve for imports—can increase its real national income by levying an optimum tariff on imports. When discrimination by country of origi is ruled out, the well-known Mill-Bickerdik tariff maximizes real national income b optimal exploitation of monopsony powe In this section, the original tariff impose by NW on imports from NE and S is as sumed to be of this type, contributing t

further increases producing little change in W. This somewhat inelegant discontinuity substantially simplifies the analysis, permitting W to be maximized with respect to I and T subject to a fixed value of N. With a less contrived form for W, the new optimum value of X would be below X^0, reflecting the now higher cost (in terms of I and T) of maintaining a given value of X.

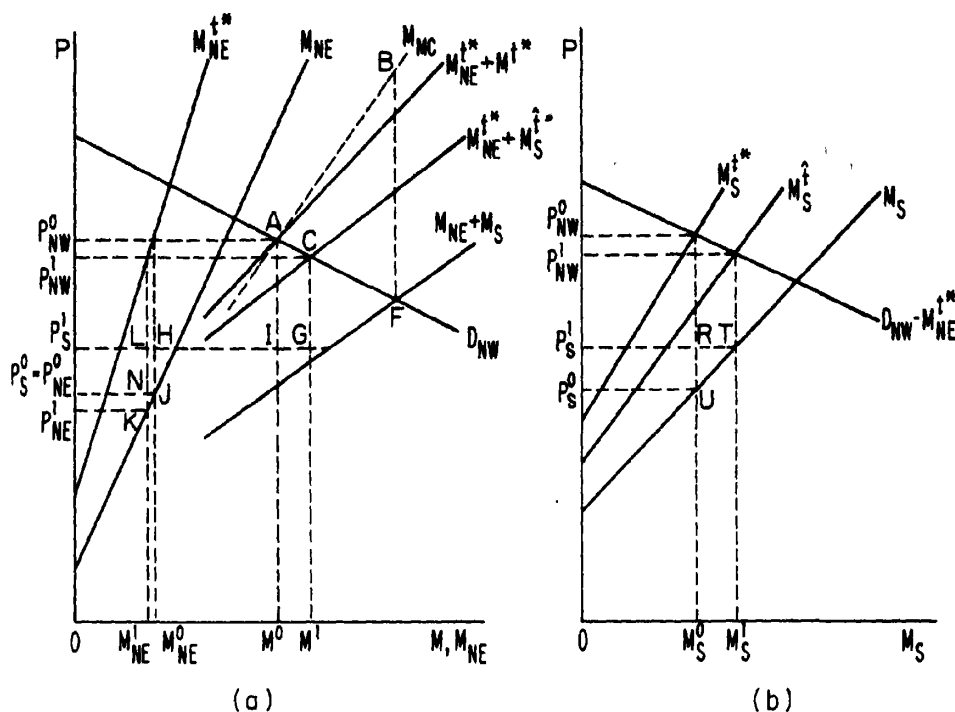



FIGURE 2

NW welfare through its effects on real national income I.

The initial absence of tariff discrimination may be explained in several ways, and the consequences of preferential trade depend in part on which of these explanations is applicable. The nondiscriminatory tariff policy of NW may indicate that discrimination does not yield any extra benefits above those obtained by the optimum nondiscriminatory tariff, as would be the case if the export supply curves of NE and S have the same elasticity at any given price. Alternatively, NW may be unable to segment the market for its imports without the collusion of at least one of the suppliers. And even when collusion between NW and one of the other trading blocs could improve the welfare of both by segmenting the market, international trading rules such as those of *GATT* may rule out this discriminatory policy.

Figure 2 illustrates the initial optimum tariff position.[19] The combined supply curve of imports from NE and S is imperfectly elastic; marginal cost to NW of imports from the two sources is shown by the M_{MC} schedule. The tariff rate t^* equal to $(P_{NW}^0 - P_S^0)/P_S^0$ maximizes the gain to NW from nondiscriminatory trade with NE and S. In the initial equilibrium, P_{NW}^0 is the domestic price in NW, while the supply price of imports is $P_S^0 = P_{NE}^0$. The gain in real income to NW through use of the optimum tariff over the free trade equilibrium (point F in Figure 2a) is measured by the area FAB.

[19] The partial equilibrium tariff discrimination model used here is based on Johnson (1962). Alexander Henderson has shown that partial equilibrium price discrimination is easily translated into general equilibrium by the use of reciprocal demand curves. This approach is followed by Kemp and extended by Caves through the explicit introduction of "purposive welfare-maximizing behavior" by a trader entering into a preferential agreement

Now NW institutes partial[20] preferences on imports from S, so that these imports are subject to a reduced tariff rate $\hat{t}$, $0 < \hat{t} < t^*$. In the new equilibrium situation (point C in Figure 2a), domestic price has fallen to P^1_{NW}, imports from NE are reduced by $M^0_{NE} - M^1_{NE}$, while NW's total imports rise by $M^1 - M^0$ as a result of increased consumption and decreased domestic output. Thus S experiences a terms of trade gain on its initial level of exports to NW (area $P^1_S RUP^0_S$ in Figure 2b) and a producer's surplus gain on its additional exports (area RTU). It follows that S enjoys an increase in real income of

$$\Delta T = P^1_S M^1_S - P^0_S M^0_S - \int_{M^0_S}^{M^1_S} P_S dM_S$$

which is measured by area $P^1_S TUP^0_S$. The net real income cost to NW of the partial tariff preference granted to S is given by

$$-\Delta I = \Delta P_S(M^0_{NE} - M^1_{NE}) + \Delta P_S M^0_S + [1/2\Delta P_{NW} - (P^1_{NW} - P^1_S)](M^1 - M^0) + \Delta P_{NE} M^1_{NE}$$

which corresponds to area

$$LHJN + P^1_S RUP^0_S - ACGI - P^0_{NE} NKP^1_{NE}$$

Real income of NE decreases through both a terms of trade loss (area $P^0_{NE}NKP^1_{NE}$) and a loss of producer's surplus (NJK).

By assumption NW's initial tariff was optimal only with respect to the rest of the world taken as a single trading unit. In the case illustrated, M_S is more elastic than M_{NE} at any common supply price. Thus, the marginal cost of imports from NE exceeds that of imports from S in the initial nondiscriminatory equilibrium. Under these circumstances, a preferential tariff on imports from S represents a second best approximation to optimal discrimination by NW between the two suppliers. There is therefore a possibility that, whatever the real income gain to S, NW may also obtain an increase in real income (a negative net cost) which is a reflection of the gains from price discrimination. Granting tariff preferences to S could thus unambiguously increase NW welfare through increases in both I and T.[21] Even if the net cost is positive but lower than the real income gain to S, it may still cost NW less than a dollar of real income foregone to transfer a dollar of resources to S. In effect, NW uses its monopsony power to tax NE for the benefit of S.

Given the *linear* supply functions depicted in Figure 2, optimal discrimination would require a lower tariff $\hat{t}_S < t^*$ on imports from S and a higher tariff $\hat{t}_{NE} > t^*$ on imports from NE, with total imports unchanged. If NW were to *raise* its MFN tariff rate to $\hat{t}_{NE}$ while allowing imports from S "preferential access" at the rate $\hat{t}_S$, optimal discrimination would be achieved and both NW and S would gain unambiguously relative to the initial equilibrium.

[20]Because of the second best nature of the problem, the case in which the tariff on preferred imports is reduced rather than completely eliminated is of particular interest.

[21]The real income of NE has not been assumed to enter the NW welfare function. If a real income loss to NE lowers NW welfare, this will provide an offset to the welfare gain from increased I and T.

REFERENCES

M. D. Bale, "Estimates of Trade—Displacement Costs for U.S. Workers," *J. Int. Econ.*, Aug. 1976, *6*, 245–50.

R. Blackhurst, "Tariff Preferences for LDC Exports: A Note on the Welfare Component of Additional Earnings," *Riv. Int. Sci. Econ. Com.*, Dec. 1971, *18*, 1180–88.

———, "General Versus Preferential Tariff Reduction for LDC Exports: An Analysis of the Welfare Effects," *Southern Econ. J.*, Jan. 1972, *40*, 350–62.

R. E. Caves, "The Economics of Reciprocity: Theory and Evidence on Bilateral Trading Arrangements," in Willy Sellekaerts, ed., *International Trade and Finance*, London 1974.

C. A. Cooper and B. F. Massell, (1965a) "Toward a General Theory of Customs Unions for Developing Countries," *J.*

Polit. Econ., Oct. 1965, *73*, 461–76.

——— and ———, (1965b) "A New Look at Customs Union Theory," *Econ. J.*, Dec. 1965, *75*, 742–47.

R. N. Cooper, "The European Community's System of Generalized Tariff References: A Critique," *J. Develop. Stud.*, July 1972, *8*, 379–94.

W. M. Corden, *Trade Policy and Economic Welfare*, Oxford 1974.

A. M. Henderson, "A Geometrical Note on Bulk Purchase," *Economica*, Feb. 1948, *15*, 61–69.

H. G. Johnson, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, *68*, 327–45.

———, "The Economic Theory of Customs Union," reprinted in *Money, Trade and Economic Growth*, Cambridge, Mass. 1962, ch. 3.

———, "Optimal Trade Intervention in the Presence of Domestic Distortions," in Richard E. Caves, et al., eds., *Trade, Growth and the Balance of Payments*, Amsterdam 1965.

———, "Trade Preferences and Developing Countries," *Lloyds Bank Rev.*, Apr. 1966, *80*, 1–18.

Murray C. Kemp, *A Contribution to the General Equilibrium Theory of Preferential Trading*, Amsterdam 1969.

M. B. Krauss, "Recent Developments in Customs Union Theory: An Interpretive Survey," *J. Econ. Lit.*, June 1972, *10*, 413–36.

Richard G. Lipsey, *The Theory of Customs Unions: A General Equilibrium Analysis*, London 1970.

S. P. Magee, "The Welfare Effects of Restrictions on *U.S.* Trade," *Brookings Papers*, Washington 1972, *3*, 645–701.

R. McCulloch, "United States Preferences: The Proposed System," *J. World Trade Law*, Mar./Apr. 1974, *8*, 216–26.

James Meade, *The Theory of Customs Unions*, Amsterdam 1955.

T. Murray, "How Helpful is the Generalized System of Preferences to Developing Countries?," *Econ. J.*, June 1973, *83*, 449–55.

J. Pinera, "World Income Redistribution Through Trade," unpublished doctoral dissertation, Harvard Univ. 1974.

Jaroslav Vanek, *General Equilibrium of International Discrimination*, Cambridge, Mass. 1965.

Jacob Viner, *The Customs Union Issue*, New York 1950.

International Monetary Fund, *IMF Survey*, June 23, 1975, *4*, 186–89.

The Regulated Firm with a Fixed Proportion Production Function

By THOMAS E. KENNEDY*

Harvey Averch and Leland Johnson in their pioneering paper on regulatory modeling found that a firm regulated by a maximum allowed rate of return on capital would generally find it advantageous to substitute capital for other inputs to produce its output in an overly capital intensive manner. However, they and others have suggested that this misallocation of inputs would not exist if the firm's production function was of the fixed proportion type so that the firm could not substitute inputs.[1]

This paper examines the behavior of the regulated firm with a fixed proportion production function. It concludes that the firm produces efficiently only if demand is sufficiently elastic so that marginal revenue exceeds marginal noncapital cost. Lowering the allowed return on capital in an attempt to increase output beyond the point where marginal revenue equals marginal noncapital cost will prompt the firm to acquire idle capital and will not bring forth any increase in output.

The L-shaped isoquant associated with fixed proportion production functions may be particularly relevant to public utilities. The complex technology utilized involves both limited substitutability of other inputs for existing plant and equipment, and long lifetimes for this capital.[2] Even if the L-shaped isoquant is not applicable to new capital investment, it may be applicable for existing capital plant and equipment. Thus a significant part of the firm's capacity can be considered to have minimal input substitution possibilities over a substantial period of time. The firm without factor substitutability can react to regulation by being overly capital intensive only by padding its rate base with idle capital.

To develop the condition under which the firm with a fixed proportion production function would produce inefficiently, the following notation will be employed:

π = profit
q = output
$R(q)$ = revenue function
K = physical units of capital
L = physical units of labor[3]
$q = q(\min(K/a, L/b))$ = the fixed proportion production function with $a > 0$ and $b > 0$
r = cost of obtaining funds, the unit cost of capital[4]

*Assistant professor of economics, Kansas State University. I am grateful to C. F. Christ, P. J. Gormeley, B. L. Jaffee, and F. T. Sparrow for their comments on an earlier draft of this paper. I also acknowledge the very useful suggestions made by the anonymous referee.

[1]They write. "If it [the production function] involves fixed proportions, . . . the regulated firm is constrained to the efficient expansion path" (p. 1057). Gordon Corey echoed their result in a recent article in which he states, "If there were fixed proportions in production, the expansion paths of the regulated and unregulated firm would be identical" (p. 364). Frederick Scherer also agrees, "In the limiting case of zero substitution elasticity (associated with L-shaped isoquants) there will be no departure from the socially optimal capital/labor ratio" (p. 532). While David McNicol (p. 432, fn. 10) questions the Averch-Johnson result that no misallocation would occur with a fixed proportion production function, he provides no conditions for its occurrence or nonoccurrence.

[2]For example, the Federal Power Commission (pp. 1–29) reports steam electric plant equipment life is estimated to about thirty to thirty-five years. Also depreciation rates for regulated industry, while admittedly conservative, suggest long lifetimes for physical capital. Alfred E. Kahn (p. 118) gives depreciation rates for regulated industries in the 2 to 5.4 percent range which implies physical capital lives in excess of fifteen years.

[3]The term labor is used here by convention, but can be thought of more generally as including all noncapital inputs.

[4]The cost of capital is equal to its acquisition cost multiplied by the cost of obtaining funds. In the following, capital units will be measured so that the acquisition cost of a unit of capital is equal to unity.

w = price per physical unit of labor
s = allowed rate of return per unit of capital with $s = r + v, v > 0$

The objective of the firm is to maximize profit:

$$(1) \qquad \pi = R(q) - rK - wL$$

subject to the constraint limiting the firm's rate of return on capital,

$$(2) \qquad R(q) - sK - wL \leq 0$$

Since we are concerned with cases in which regulation does affect the firm, henceforth we assume the equality form of the constraint. Combinations of capital and labor which satisfy the equality form of (2) form the iso rate-of-return curve for a given s. Following William Baumol and Alvin Klevorick, we define the locus of maximum profit points for alternative iso rate-of-return curves as the expansion path of the firm. We are concerned with finding this path and comparing it to the efficient expansion path with fixed proportion production, $K/a = L/b$.

To find the regulated firm's expansion path, note that since q depends on the minimum of K/a and L/b, the iso rate-of-return constraint can be written as

$$(3) \qquad R(q(L)) - wL - sK = 0 \text{ for } \frac{K}{a} \geq \frac{L}{b}$$

$$(4) \qquad R(q(K)) - wL - sK = 0 \text{ for } \frac{K}{a} \leq \frac{L}{b}$$

Since allowed profits can be increased only by increasing capital, equation (3) which allows for over-capital intensity is of prime interest and will be considered here.[5]

Because the regulated profit is equal to $(s - r)K$, the firm maximizes its profit by oosing that point on the relevant iso rate-of-return curve which has the maximum capital. The firm would operate in the overly capital intensive region if the slope of the iso rate-of-return contour becomes vertical. Differentiating (3) we obtain this condition:

$$(5) \qquad \frac{dK}{dL} = \frac{1}{s}\left(\frac{dR}{dL} - w\right) = 0$$

The condition for the firm to be overly capital intensive is that the marginal revenue product of labor along the relevant iso rate-of-return curve is equal to the price of labor. Alternatively, if the expression in parentheses is divided by the marginal product of labor, the condition can be stated in terms of output as marginal revenue is equal to marginal labor cost. When this condition occurs it becomes profitable for the firm to pad its rate base with excess capital rather than continuing to expand output to meet the regulatory constraint.

In the overly capital intensive range, revenues depend only on L. Assuming that dR/dL is a monotonically declining function of L, equation (5) will have only one root.[6] This root represents the maximum amount of L (denoted by L^*) which the firm would employ and results in the maximum output the firm would produce.

While the firm may be inefficient, it can be shown that there does exist a range of efficient production along the firm's expansion path in which regulation can decrease the firm's profits and increase its output while maintaining efficient production. The unconstrained firm chooses $K/a = L/b$ so that profit can be written as

$$(6) \qquad \pi = (R(L)) - wL - r\frac{a}{b}L$$

and reaches a maximum profit at the root of

$$(7) \qquad \frac{dR}{dL} - w - r\frac{a}{b} = 0$$

erefore, the cost of capital is equal to r. This produre eliminates the need for a separate term for the quisition cost of capital without loss of generality. is simplification was also made by Averch-Johnson.

[5]For another approach to the problem including a oof that the firm would not be overly labor intene, see the author.

[6]Since dR/dq will be declining, dR/dL will also decline except where economies of scale are so great that the rise in the marginal product of L is able to offset the fall in marginal revenue. At the chosen operating point of the firm, dR/dL will be declining even if dR/DL is not monotonic.

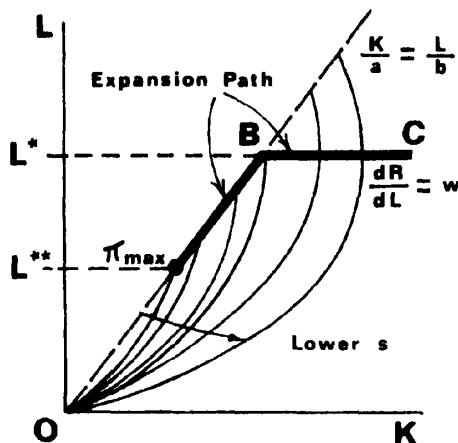



FIGURE 1

Denote this root by L^{**}. Since $w + (a/b) \cdot r > w$, $L^{**} < L^{*}$, and a range of efficient production exists between L^{**} and L^{*}.

The above results can be summarized in Figure 1 which illustrates the iso rate-of-return contours for the overly capital intensive range of input use. The firm's expansion path is $\pi_{max} BC$. For rates of return between the unconstrained profit-maximizing rate of return and the iso rate-of-return curve passing through point B where $dR/dL = w$, regulation is effective in decreasing profit and increasing output while maintaining efficient production. However, if the regulators attempt to lower the allowed return below the level attained at point B, the result will be no change in output while profit is lowered as the firm acquires enough excess capital to achieve the lower allowed rate of return along range BC.

The results found here show that lack of substitutability of inputs does not insure economically efficient use of inputs for the rate-of-return regulated firm. There can be significant differences between the minimum cost required for production of a given output and the cost in terms of resources used by the regulated firm. Regulators need to be cognizant of the possibility of inefficient behavior on the part of the regulated firm even when input substitution opportunities do not exist for the firm.

The statement of Averch-Johnson that the regulated firm with a fixed proportion production function will produce efficiently is correct only so long as marginal revenue is greater than marginal labor (noncapital) cost. When regulation is such that with efficient production marginal revenue would not cover marginal labor cost, the regulated firm would maximize its profits by producing at the point at which marginal revenue equals marginal labor cost and expanding its rate base with idle capital sufficiently to meet the rate-of-return constraint.

In a sense, the misallocation of resources in the fixed proportion case may be of a worse nature than in the substitutable input case. The misallocation associated with substitutable inputs is analogous to the firm using the wrong input prices in its input decision. In the fixed proportion case, no positive input prices could cause the use the resources which the regulated fir chooses. "Second best" consideratio might possibly make it socially desirable f the regulated firm to be overly capital i tensive in terms of market input pric However, such considerations would nev call for the outright waste of scarce r sources which may occur in the case of t regulated firm with a fixed proportion pr duction function. Thus, while it is true th the fact that a regulated firm's productı function is of the fixed proportion type m

prevent inefficiency given sufficiently elastic demand for the product, this same type of production function may result in the outright waste of inputs if demand is not sufficiently elastic.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

W. J. Baumol and A. K. Klevorick, "Input Choices and Rate of Return Regulation: An Overview of the Discussion," *Bell J. Econ.*, Autumn 1970, *1*, 162–90.

G. R. Corey, "The Averch and Johnson Proposition: A Critical Analysis," *Bell J. Econ.*, Spring 1971, *2*, 358–73.

Alfred E. Kahn, *The Economics of Regulation: Principles and Institutions*, Vol. 2, New York 1971.

T. E. Kennedy, "Regulatory Control of Monopoly: Limitations, Methods and Effects," unpublished doctoral dissertation, Johns Hopkins Univ. 1975.

D. L. McNicol, "The Comparative Static Properties of the Theory of the Regulated Firm," *Bell J. Econ.*, Autumn 1973, *4*, 428–53.

Frederick M. Scherer, *Industrial Market Structure and Economic Performance*, New York 1971.

Federal Power Commission, *The 1970 National Power Survey*, Part IV, Washington.

Intergenerational Equity and the Investing of Rents from Exhaustible Resources

By JOHN M. HARTWICK*

Invest all profits or rents from exhaustible resources in reproducible capital such as machines. This injunction seems to solve the ethical problem of the current generation shortchanging future generations by "overconsuming" the current product, partly ascribable to current use of exhaustible resources.[1] Under such a program, the current generation converts exhaustible resources into machines and "lives off" current flows from machines and labor. Under such a program one might assume that in some sense the total stock of productive capital was never depleted since ultimately the exhaustible resource stock will be transmuted into a stock of machines and, given that machines are assumed not to depreciate, no stock either of machines or of exhaustible resources is ever consumed. If in this sense the stock of productive capital is not being depleted, what can one say about the time path of current output and current consumption per head? For the case of per capita consumption remaining constant over time, one could say that no generation was better off than another. Intergenerational equity was being achieved.[2] For simplicity, we shall assume *ZPG* or a constant population so we need only ask what happens to the time path of aggregate consumption. Let me restate the problem in brief: if society invests all rents from exhaustible resources in reproducible capital goods, and invests only this amount, i.e., consumes the remainder of the product given population constant, will consumption and output rise, remain constant, or fall over time?

I shall formally set this problem out below and solve it for the case of a Cobb-Douglas technology. The Cobb-Douglas technology has the important property that each input (in particular, the flow of minerals from an exhaustible resource) is essential for producing a positive output of the single produced commodity. Thus the economy cannot exhaust any natural resource and continue to have positive consumption and output. Beckmann (1974, 1975), Solow, and Solow and Wan have used the Cobb-Douglas technology in their analyses of utilization of exhaustible resources in aggregate dynamic models.

Production in the model at period t will be assumed to require inputs of reproducible capital $k(t)$, flows of mineral from a exhaustible resource $y(t)$ and labor. Th labor force is constant so we can set it t one unit. The $k(t)$, $y(t)$, commodity outpu $x(t)$, and consumption $c(t)$ are defined i per capita terms. The technology $f(k(t$ $y(t), 1)$ will be assumed to exhibit constan returns to scale so that $f(\cdot)$ is homoge

*Associate professor, Queen's University

[1]The idea for this paper arose after hearing a seminar by Anthony Scott on resource policy. He estimated the returns Canadians might receive in 1975 if they had invested all resource royalties in assets yielding the current rates of interest prevailing since 1911. The interest each year was to be consumed He labeled such a strategy as a "Saudi Arabian" program!

[2]See Kenneth Arrow for a systematic exploration of savings rules and intergenerational equity within the context of a model of accumulation of reproducible capital in the absence of exhaustible resources. The current interest in intergenerational equity was aroused by remarks of John Rawls (sec. 44). One should of course consult Rawls, Arrow, an Robert Solow for an introduction to the diverse nc tions of intergenerational equity which have bee proposed for consideration in the current investiga tions. Rawls was concerned with the problem o balancing the relative burden of savings on earl generations with the burden on later generations Capital was being *accumulated* over some part of society's program of consumption and investment With exhaustible resources, one must be concerne with *forestalling decumulation* of society's produc tive capital in order to achieve some notion of inter generational equity.

neous of degree one. The value of $x(t) = f(\cdot)$ will be zero if any argument of $f(\cdot)$ is zero. That is, each input is essential. The marginal productivities $\partial f/\partial k$ and $\partial f/\partial y$ are assumed to be positive; $\partial^2 f/\partial k^2$ and $\partial^2 f/\partial y^2$ are assumed to be negative. Let $f_k \triangleq \partial f/\partial k$, $f_y \triangleq \partial f/\partial y$, $f_{kk} \triangleq \partial^2 f/\partial k^2$, $f_{ky} \triangleq \partial^2 f/\partial k \partial y$, and $f_{yy} \triangleq \partial^2 f/\partial y^2$. A D before a variable will indicate the time derivative of that variable (for example, $Dk \triangleq dk/dt$). At any instant of time, the product $x(t)$ is completely divided between current consumption $c(t)$, investment Dk and extraction costs $ay(t)$, where a is the cost measured in units of the single produced commodity of extracting one unit of the exhaustible resource. Thus, we have our accounting relation

$$x(t) = c(t) + Dk + ay(t)$$

Our savings or investment function is

$$Dk = (f_y - a)y(t) \tag{1}$$

Efficiency of exhaustible resource extraction requires that the rate of return from a unit of reproducible capital equal the rate of return from owning a unit of deposits of the exhaustible resource.[3] In price terms, this condition is characterized by the current capital gain on mineral deposits being equal to the interest rate or rate of return on reproducible capital. In our one-commodity world, this condition is satisfied by *the rate of change in the marginal product of the mineral being equal to the marginal product of reproducible capital.* This is sometimes referred to as the Hotelling Rule. It characterizes the efficient exploitation of an exhaustible resource. That is

$$\frac{d \log (f_y - a)}{dt} = f_k \tag{2}$$

or

$$f_{yy}Dy + f_{yk}Dk = f_k(f_y - a) \tag{2'}$$

Relations (1) and (2) define the dynamics of the economy. There are two differential equations in the variables $y(t)$ and $k(t)$. We require initial values $k(0)$ and $y(0)$ in order to define the time paths of $y(t)$ and $k(t)$. We shall assume that $k(0)$ and $y(0)$ are selected so that the initial stock of exhaustible resource S is precisely sufficient to sustain the economy over infinite time. We shall remark below that there exists a finite S which will yield the consumption path below. By definition, $dS/dt = -y(t)$, where the stock S is defined in per capita terms.

Aggregate output is rising, constant, or falling over an interval of time as $Dx \gtreqless 0$. Now from the definition of the production function, we get

$$Dx = f_k Dk + f_y Dy \tag{3}$$

For the case of the Cobb-Douglas technology, we have

$$x = k^{\alpha} y^{\beta} 1^{\gamma}$$

with $\alpha + \beta = 1$ and $f_k \triangleq \alpha x/k$ and $f_y \triangleq \beta x/y$. Also $f_{yy} \triangleq \beta x(\beta - 1)/y^2$ and $f_{yk} \triangleq \alpha\beta x/yk$.

For the case of the Cobb-Douglas technology, (2′) becomes

$$f_y Dy - xDy/y + f_k Dk = (y/\beta) f_k (f_y - a)$$

and substituting for Dk from (1), we get

$$\beta[f_y Dy + f_k(f_y - a)y] = f_y Dy + f_k(f_y - a)y \tag{4}$$

Since $0 < \beta < 1$, equation (4) can only be satisfied if $f_y Dy + f_k(f_y - a)y = 0$ but $f_y Dy + f_k(f_y - a)y$ is the right-hand side of (3). Thus we have established that x will be constant over time and since $c(t) = (1 - \beta)x(t)$, (recall $f_y y = \beta x$), we have the result that consumption per head will be constant over time. Given the finiteness of natural resource stock, it will be necessary over infinite time to have the current flow of resources extracted asymptotically approach zero as time tends to infinity. By Solow's definition of intergenerational equity—namely per capita consumption remaining constant over time—we have established that the savings investment rule (invest all net returns from exhaustible re-

[3] Since in a well-behaved problem (e.g., the case with a Cobb-Douglas production function) f_y always increases as $t \to \infty$, one has only to assure that the extraction costs are such that $(f_y - a) > 0$ at t_0.

sources in reproducible capital) implies intergenerational equity. A perusal of the mathematics of Solow's paper indicates that this result was implicit in his mathematics—to preserve $Dc = 0$, society should invest the current returns from the utilization of flows from the stock of exhaustible resources.

We have in fact obtained the rule for a model with nonzero extraction costs. Solow had no extraction costs in his formulation. He proved that the existence of a solution required that $\beta < \alpha$. We take this as a necessary condition for existence. It implies that the share of output ascribable to natural resources be less than that share ascribable to reproducible capital—a condition which empirical results indicate is unambiguously satisfied. To be precise, the only model in which the existence of a solution with c positive over infinite time and S finite has been established is the above Cobb-Douglas case with extraction costs set at zero. On reading the above note Solow pointed out that the rule "invest exhaustible resource rents and c will remain constant" is very general. To see this substitute from (1) and (2) in the relation $Dx = f_k Dk + f_y Dy$ for f_k and Dk to get

$$Dx = \left\{ \frac{df_y}{dt} \Big/ (f_y - a) \right\} (f_y - a) y + f_y Dy$$

$$= \frac{d(f_y y)}{dt} = \frac{d(Dk + ay)}{dt}$$

Given $x = c + Dk + ay$ we conclude that $Dc = 0$ regardless of whether $Dx = 0$. Thus we have established, for general technologies, the rule: "the investment of current exhaustible resource returns in reproducible capital implies per capita consumption constant." For the Cobb-Douglas case $Dk + ay = f_y y = \beta x$. Thus from above we have $Dx = \beta Dx$ and since $\beta \neq 1$, $Dx = 0$. If there is depreciation of reproducible capital at the rate δ per unit capital per unit time, then net capital accumulation is currently of the amount $dk/dt + \delta k(t)$ and given our savings rule, equation (1) becomes $Dk + \delta k = (f_y - a) y(t)$. Reworking the steps for solving for Dc above reveals that $Dc = -\delta f_k k$. Hence our savings investment rule will not provide for the maintaining of per capita consumption constant over time. The current decline in per capita consumption is simply the amount of the produced commodity required to offset the current amount of depreciation in the reproducible capital. Arrow's results would not turn on whether he has reproducible capital depreciate; he does not explicitly treat depreciation. In my 1976 paper the present model is extended to cover cases of many exhaustible resources. The intergenerational equity result has also been established in a Uzawa two-sector model with an exhaustible resource.

REFERENCES

K. Arrow, "Rawls' Principle of Just Saving," *Swedish J. Econ.*, Dec. 1973, *75*, 323–35.

M. J. Beckmann, "A Note on the Optimal Rate of Resource Exhaustion," *Rev Econ. Stud.* Symposium, 1974, 121–22.

———, "The Limits to Growth in a Neoclassical World," *Amer. Econ. Rev.*, Sept. 1975, *65*, 695–99.

J. M. Hartwick, "Substitution Among Exhaustible Resources and Intergenerational Equity," *Rev. Econ. Stud.*, forthcoming.

John Rawls, *A Theory of Justice*, Cambridge, Mass. 1971.

R. M. Solow, "Intergenerational Equity and Exhaustible Resources" *Rev. Econ. Stud.* Symposium, 1974, 29–46.

——— **and F. Y. Wan**, "Extraction Costs the Theory of Exhaustible Resources *Bell J. Econ*, Autumn 1976, *7*, 359–70.

Tariffs vs. Quotas as Revenue Raising Devices under Uncertainty

By Partha Dasgupta and Joseph Stiglitz*

The relative merits of a tariff and a quota at the border for achieving a government objective have been discussed a good deal in the literature.[1] It has long been recognized that, provided the government auctions off the quota, the optimum pure tariff and the optimum pure quota are equivalent in a competitive world with no uncertainty.[2] The proposition continues to hold if there is uncertainty, but where every agent, including the government, can monitor (and therefore distinguish) costlessly every state of nature. However, in such a world the equivalence is between a pure tariff and a pure quota that are both functions of the state of nature.[3] It is this last observation that leads us to suspect that the equivalence result is of rather limited practical use. One would imagine that the possible states of nature are large in number. It is then difficult to envisage a government announcing a trade policy that is contingent entirely on the state of nature. Such a policy would be costly to calculate and difficult to comprehend. We are therefore encouraged to simplify a good deal and to restrict the set of admissible trade policies. But this is a difficult matter. It is by no means immediate what restrictions would appear as being natural to contemplate. The border policies that are most commonly resorted to by governments are fixed tariff rates and fixed quantity restrictions. They are often regarded as polar forms of trade restrictions (one involving prices and the other involving quantities). They have very different effects: a fixed tariff on a commodity stabilizes its domestic price in the face of random domestic demand and supply but a fixed international price; while a quota stabilizes its domestic price if its international price is random but its domestic demand and supply functions are fixed.

The major purpose of this paper is to examine the relative merits of these two trade policies in the presence of uncertainty. The central result that we shall present here came somewhat as a surprise to us. Under the conventional criterion of maximizing the expected value of net consumer's surplus, the optimum fixed tariff is superior to the optimum fixed quota. The result continues to hold if instead the maxi-min criterion is followed. Section I is concerned with this issue. In Section II we discuss in what sense both tariffs and quotas may be viewed as special cases of a more general class of trade policies, and how the problem of the choice of an optimum trade policy may be viewed as an example of a general class of problems arising out of imperfect information.

I. Tariffs versus Quotas

Imagine a commodity that can be both produced domestically under competitive conditions and at the same time imported. Denoting by q its domestic price and by D its domestic demand, we suppose that the market demand curve can be represented as[4]

$$q = \tilde{\alpha} - \beta D \tag{1}$$

where β is a positive constant and where $\tilde{\alpha}$

*Reader in economics, London School of Economics, and professor of economics, University of Oxford, respectively. This work was supported by National Science Foundation Grant SOC74-22182 at Stanford University. We are most grateful to Phillip Kott for valuable comments.

[1] See, for example, Jagdish Bhagwati.

[2] By a pure tariff rate we mean a tariff rate that is independent of the quantity traded. Likewise, by a pure quota we mean a single quantity restriction.

[3] This will be demonstrated formally in the example that follows.

[4] In what follows, a tilde above a variable will imply that the variable is random, and a bar above it will denote its expected value. Thus, for example, $E(\tilde{\alpha}) \equiv \bar{\alpha}$.

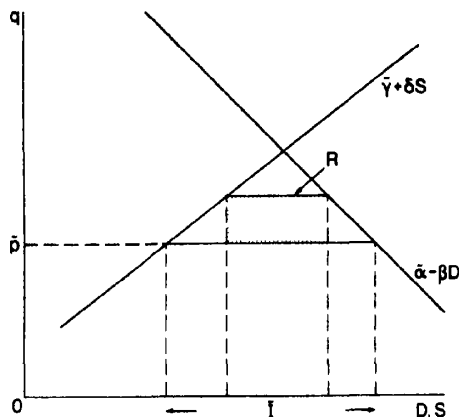



FIGURE 1

is a random variable with a known distribution. Writing S as the domestic supply of the commodity, we take it that the domestic supply function can be represented as[5]

$$q = \tilde{\gamma} + \delta S \tag{2}$$

where δ is a positive constant and where $\tilde{\gamma}$ is a random variable. In short, we are postulating linear demand and supply curves, each of which possesses an unbiased shift parameter. (See Figure 1.) Alternatively, one could suppose that all the random variables have small variances, so that we would be justified in taking linear approximations of the domestic demand and supply functions at the optimum tariff point.

The economy in question is assumed to be small in that its import requirement does not influence the foreign price. The import price $\tilde{p}$ is not known with certainty but is random with a known distribution.

Given that the commodity is domestically produced under competitive conditions the domestic cost function $C(S)$ is the integral of supply curve:

$$C(S) = \tilde{\gamma}S + \frac{\delta}{2}S^2 = (\tilde{q}^2 - \tilde{\gamma}^2)/2\delta \tag{3}$$

(Without loss of generality we are setting the constant of integration at zero.) Likewise, consumer's gross benefit $B(D)$ from (1) can be expressed as

$$B(D) = \tilde{\alpha}D - \frac{\beta}{2}D^2 = (\tilde{\alpha}^2 - \tilde{q}^2)/2\beta \tag{4}$$

Given that we are concerned here with the relative merits of a pure tariff and a pure quota we need an account of the rationale for introducing a trade restriction. Assume then that the government desires to introduce such a restriction with a view to raising a given expected level of revenue R, and to keep matters simple we take it that the government is risk neutral. Now it is plain that there are various manners in which the government can introduce trade restrictions in order to ensure an expected level of revenue R. Let E denote the expectation operator. We suppose that it ranks various feasible policies in accordance with the function

$$W = E(B(\tilde{D}) - C(\tilde{S}) - \tilde{p}\tilde{I})$$

where $\tilde{I}$ is the equilibrium level of import.[6]

By way of contrast let us look at the first best formulation of the problem initially. For this economy a state of nature is characterized by a triplet of numbers $(\tilde{\alpha}, \tilde{\gamma}, \tilde{p})$. In order to raise the expected level of revenue R, the government would announce an *ad valorem* tariff t, contingent on the state of nature. Writing by $I(\tilde{\alpha}, \tilde{\gamma}, \tilde{p})$ the quantity imported in the state of nature $(\tilde{\alpha}, \tilde{\gamma}, \tilde{p})$, and using (1) and (2), a market equilibrium would be characterized by the condition

$$I(\tilde{\alpha}, \tilde{\gamma}, \tilde{p}) = \frac{\tilde{\alpha}}{\beta} + \frac{\tilde{\gamma}}{\delta} - \tilde{p}(1 + t(\tilde{\alpha}, \tilde{\gamma}, \tilde{p}))\frac{(\beta + \delta)}{\beta\delta} \tag{5}$$

In what follows we shall take it for sim-

[5]It is, of course, possible to present the analysis by postulating directly the excess demand function, rather than working separately with the demand and supply functions. The final result that we shall present subsequently is, however, easier to dissect if we consider them separately.

[6]In other words, the government ranks policies in accordance to their contributions to the expected value of the sum of the surpluses accruing to consumers and producers. (Note that it makes no difference whether the government ranks projects according to $E[\tilde{B} - \tilde{C} - p\tilde{I}]$ or $E[\tilde{B} - \tilde{C} - \tilde{q}\tilde{I}]$ since they differ by exac R.)

plicity that the ranges of the random variables $\tilde{\alpha}$, $\tilde{\gamma}$, and $\tilde{p}$ are small in the sense that in the absence of any trade restrictions (i.e., $t(\tilde{\alpha},\tilde{\gamma},\tilde{p}) = 0$), one has for all realizations of $\tilde{\alpha}, \tilde{\gamma}$, and $\tilde{p}$, $I(\tilde{\alpha},\tilde{\gamma},\tilde{p}) > 0$. To keep the analysis uncomplicated we shall subsequently assume as well that R is small in a sense that will be made precise.

The government's problem would then consist of determining a tariff schedule $t(\tilde{\alpha},\tilde{\gamma},\tilde{p})$ that will maximize $E(\tilde{B}(D) - \tilde{C}(S) - \tilde{p}\tilde{I})$, subject to the constraint (5) and the condition

$$E(\tilde{p}t(\tilde{\alpha},\tilde{\gamma},\tilde{p})I(\tilde{\alpha},\tilde{\gamma},\tilde{p})) = R \tag{6}$$

Assume for the moment that an optimum exists. It is of course plain that whether the government announced the resulting optimal tariff schedule $t^*(\tilde{\alpha},\tilde{\gamma},\tilde{p})$ or instead uses (5) to auction off the corresponding import quota schedule $I^*(\tilde{\alpha},\tilde{\gamma},\tilde{p})$ is a matter of indifference. This is the classical equivalence between tariffs and quotas.

More generally, from equations (3), (4), and (5), we can express expected net benefits as

$$W = \frac{E(\tilde{\alpha}^2)}{2\beta} + \frac{E(\tilde{\gamma}^2)}{2\delta} - \frac{\beta+\delta}{\beta\delta}\left[\frac{E(\tilde{q}^2)}{2} - E(\tilde{p}\tilde{q})\right] - \left[\frac{E(\tilde{\alpha}\tilde{p})}{\beta} + \frac{E(\tilde{\gamma}\tilde{p})}{\delta}\right] \tag{7}$$

It follows that policies will be ranked simply on the basis of the value of

$$Z \equiv E(\tilde{p}\tilde{q}) - E(\tilde{q})^2/2 \tag{8}$$

We are now concerned with the second best problem, where the admissible set of trade policies is severely restricted. We suppose that the government can costlessly monitor the total volume of imports, $\tilde{I}$, and can therefore base its trade restriction on $\tilde{I}$; but that $\tilde{\alpha}$, $\tilde{\gamma}$, and $\tilde{p}$ are separately unobservable, and therefore it cannot base trade policies on them.[7] In this section we focus attention on two of the simplest of such trade policies, namely: 1) a pure tariff $t(I) = t$ (a constant); and 2) a pure quota I, which is equivalent to an implicit specific tariff τ given by the rule $\tau = \tilde{q} - \tilde{p}$ for $\tilde{q} > \tilde{p}$ and $I(\tilde{\alpha},\tilde{\gamma},\tilde{p}) \leq I$, and $\tau = \infty$ for $I(\tilde{\alpha},\tilde{\gamma},\tilde{p}) > I$.[8] In what follows we analyze the two sequentially. (See the accompanying figure.)

A. *Pure Tariff*

Denote by t the pure ad valorem tariff. It follows from equation (5) that the import function is of the form

$$I = \frac{\tilde{\alpha}}{\beta} + \frac{\tilde{\gamma}}{\delta} - \tilde{p}(1+t)\frac{(\beta+\delta)}{\beta\delta} \tag{9}$$

and from the revenue constraint (6) that

$$R = E(\tilde{p}t\tilde{I}) = tE(\tilde{p}\tilde{I}) \tag{10}$$

To analyze the issues involved it is simplest (though not essential) to suppose that the random variables $\tilde{\alpha}$, $\tilde{\gamma}$, $\tilde{p}$, are all pair-wise independent of one another. We can then use equation (9) in equation (10) to obtain a quadratic equation in t:

$$t^2 - \left[\frac{\overline{p}(\overline{\alpha}\delta + \overline{\gamma}\beta)}{(\beta+\delta)E(\tilde{p}^2)} - 1\right]t + \frac{\beta\delta R}{(\beta+\delta)E(\tilde{p}^2)} = 0 \tag{11}$$

If R is too large, there will be no real solution for t. There is then no feasible pure tariff policy. Consequently we take it that R is small. Of the two real solutions of (11) it is the smaller tariff rate which yields a higher level of expected net benefits. It is this smaller value we are interested in. Denote it by t^*. To get a tidy expression for t^* it will be convenient to assume that R is small enough so as to enable one to ignore

[7]Thus we rule out by assumption the possibility of smuggling. This raises rather different issues.

[8]In what follows we shall suppose that the ranges of $\tilde{\alpha}$, $\tilde{\gamma}$, and $\tilde{p}$ are sufficiently small and the quota is not overly large. Consequently we shall take it that $\tilde{q} > \tilde{p}$ and hence the quota is binding. It follows that strictly speaking one does not need to set $\tau = \infty$ for $I(\tilde{\alpha},\tilde{\gamma},\tilde{p}) > I$. A large enough τ will do. Notice that as the quota is, by assumption, auctioned off, the implicit tariff associated with a quota is random.

all the second and higher powers of R. Consequently from (11) one has[9]

$$(12)\quad t^* \simeq \frac{\beta\delta R}{\bar{p}(\bar{\alpha}\delta + \bar{\gamma}\beta) - (\beta + \delta)E(\tilde{p}^2)}$$

It follows then that

$$(13)\quad \tilde{q} \cong \tilde{p}\left(1 + \frac{\beta\delta R}{\bar{p}(\bar{\alpha}\delta + \bar{\gamma}\beta) - (\beta + \delta)E(\tilde{p}^2)}\right)$$

Under a pure tariff scheme the market-clearing price is random as long as the foreign price is random.

Thus, on using equation (8) one obtains (for small R)

$$(14)\quad Z_{tariff} = E(\tilde{p}^2)/2$$

B. *Pure Quota*

If a pure import quota I is announced and auctioned off, the resulting domestic price q in equilibrium is obtained from equation (5):

$$(15)\quad \tilde{q} = \frac{\beta\delta}{\beta + \delta}\left(\frac{\tilde{\alpha}}{\beta} + \frac{\tilde{\gamma}}{\delta} - I\right)$$

Thus, the uncertainty in the equilibrium price resulting from a quota is due solely to the uncertainty in the domestic supply and demand conditions. The uncertainty in the foreign price has no bearing on this. All this is, of course, obvious. Now if $\tilde{q}$ is the equilibrium market price, the implicit *specific* tariff $\tilde{\tau}$ due to the quota in equilibrium is

$$(16)\quad \tilde{\tau} = \tilde{q} - \tilde{p}$$

which, on using in (15) yields

$$(17)\quad \tilde{\tau} = \frac{\tilde{\alpha}\delta + \tilde{\gamma}\beta}{\beta + \delta} - \tilde{p} - \frac{\beta\delta I}{\beta + \delta}$$

$E(\tilde{\tau}I)$ is the expected revenue by an auction of the quota I in a risk-neutral market. Therefore condition (6) implies that

$$(18)\quad R = E(\tilde{\tau}I) = IE(\tilde{\tau})$$

[9]Notice that the validity of the approximation for t^* depends on the magnitude of the right-hand side of equation (12), while the requirement that $\tilde{q} > p$ (see fn. 8) implies restriction of the *range* of $(\tilde{\alpha}, \tilde{\gamma}, \tilde{p})$.

Using equation (17) and (18) yields a quadratic expression in I:

$$(19)\quad I^2 - \left(\frac{\bar{\alpha}\delta + \bar{\gamma}\beta - \bar{p}(\beta + \delta)}{\beta\delta}\right) I + \frac{R(\beta + \delta)}{\beta\delta} = 0$$

There are then two feasible quota specifications. Plainly the larger of the two solutions of equation (19) yields higher expected net benefits: denote it by I^*. It follows that

$$(20)\quad I^* \simeq \frac{\bar{\alpha}\delta + \bar{\gamma}\beta - \bar{p}(\beta + \delta)}{\beta\delta} - \frac{R(\beta + \delta)}{\bar{\alpha}\delta + \bar{\gamma}\beta - \bar{p}(\beta + \delta)}$$

Using equations (15) and (20) now yields the equilibrium price under the pure quota scheme as

$$(21)\quad \tilde{q} = \bar{p} + \frac{R\beta\delta}{\bar{\alpha}\delta + \bar{\gamma}\beta - \bar{p}(\beta + \delta)} + \frac{(\tilde{\alpha} - \bar{\alpha})\delta + (\tilde{\gamma} - \bar{\gamma})\beta}{\beta + \delta}$$

As before, we are concerned with the level of expected net benefits under the optimum pure quota level I^*.

From equation (8) and (21) one then obtains

$$(22)\quad Z_{quota} = (\bar{p}^2 - (\sigma_\alpha^2\delta^2 + \sigma_\gamma^2\beta^2)/(\beta + \delta)^2) \div 2$$

where σ_α^2 and σ_γ^2 are the variances of $\tilde{\alpha}$ and $\tilde{\gamma}$, respectively.

C. *Comparison*

We have finally to compare (14) and (22) to determine which of the two schemes is superior. Since $\sigma_p^2 = E(\tilde{p}^2) - \bar{p}^2$, it follows that

$$(23)\quad W_{tariff} - W_{quota} = \frac{\beta + \delta}{\beta\delta}[Z_{tariff} - Z_{quota}]$$

$$= \frac{\delta\sigma_\alpha^2}{2\beta(\beta + \delta)} + \frac{\beta\sigma_\gamma^2}{2\delta(\beta + \delta)} + \frac{(\beta + \delta)\cdots}{2\beta\delta}$$

Equation (23) is the basic result of th

paper and in what follows we comment on it. Notice first that the reason why moments of order higher than the variance do not appear in (23) is the fact that the government's objective function is quadratic. Notice as well that if $\sigma_p^2 = \sigma_\alpha^2 = \sigma_\gamma^2 = 0$, then $W_{tariff} = W_{quota}$. This last is, of course, the classical equivalence result. However, the interesting feature of equation (23) is that so long as at least one of the variances is positive, $W_{tariff} > W_{quota}$. In other words, *a pure tariff is unambiguously superior to a pure quota in generating a given expected level of government revenue.*

We had not anticipated this result. Indeed, we had supposed that the relative merits of a tariff and a quota would depend on the relative steepnesses of the demand and supply functions, and possibly also on the relative magnitudes of the coefficients of variation of the different random variables. No doubt under more general formulations of the excess demand function there are circumstances in which a pure quota is a superior policy measure to a pure tariff. But a linear excess demand function is the simplest laboratory in which to raise this question. At any rate, within the confines of such a formulation the answer emerges as being unambiguous.

The result is particularly telling for the situation where $\sigma_p^2 > \sigma_\alpha^2 = \sigma_\gamma^2 = 0$. It is under this circumstance that a pure quota is a stabilizing policy. Under such a regime there is no uncertainty in the domestic equilibrium price (see equation (21)).[10] Consequently there is no uncertainty in the sum of the surpluses accruing to producers and consumers. However, with a pure tariff the domestic equilibrium price is random (see equation (13)). It might then be thought that given risk aversion on the part of the private sector, a quota would be a superior policy measure to a tariff in generating the required government revenue. What this argument overlooks is the fact that the quota does not allow for variations in imports, say, when the social cost of imports is low (because $\tilde{p}$ is low) or when the value of imports is high (for example, because $\tilde{\alpha}$ is large). This relative adaptability of tariffs has long been argued as one of its advantages. The result here makes precise the sense in which this is true.

Remarkably enough the result does not depend on the assumption of risk neutrality with respect to net consumer's surplus. Suppose instead infinite risk aversion and consequently that policies are ranked by the function

$$\hat{W} \equiv \min(\tilde{B} - \tilde{C} - \tilde{p}\tilde{I})$$

With the max-min criteria the same result obtains, for on using equations (12) and (20), routine calculations yield

$$\hat{W}_{tariff} - \hat{W}_{quota} = \frac{\beta+\delta}{2\beta\delta}\left[p_{max} - \bar{p} - \frac{R\beta\delta}{\bar{\alpha}\delta + \bar{\gamma}\beta - \bar{p}(\beta+\delta)} - \frac{(\alpha_{max} - \bar{\alpha})\delta + (\gamma_{max} - \bar{\gamma})\beta}{\beta+\delta}\right]^2 > 0$$

II. Second Best Optimum Tariff Schedules

The analysis of the previous section could be viewed as one concerning the optimum tariff schedule when the admissible set of policies is restricted to the fixed tariff and the fixed quota. It was noted that this was an immense restriction. It emerged that even for small expected revenue requirements there are only four feasible policies in all (see equations (11) and (19)). Computing the optimum border policy was then an easy enough matter. The central result of this paper was that the fixed ad valorem tariff t^* in equation (12) is the optimum of this restricted set of policies.

Now, there is of course no reason why in principle we should restrict our attention to the fixed tariff and quota as admissible policies. The tariff schedules need to be based on observables, and ought to be administratively simple. For instance, it may be relatively easy to monitor the volume of imports but difficult to monitor the true

[10]This may well be an implicit argument in justifying the European Economic Community (EEC) policy of stabilizing domestic price by setting a tariff contingent on the imported price of commodities which is precisely what a quota achieves

price (for example, because of kickbacks, special credit facilities, etc.). Thus we could imagine the government announcing an ad valorem tariff rate $t(I)$. In order that such a function yields an outcome it must result in a real valued random variable $\tilde{I}$, satisfying the conditions

$$\tilde{I} = \frac{\tilde{\alpha}}{\beta} + \frac{\tilde{\gamma}}{\delta} - \tilde{p}(1 + t(\tilde{I}))(\beta + \delta)/\beta\delta$$

and

$$R = E(\tilde{p}t(\tilde{I})\tilde{I})$$

The aim may then be to select a schedule $t^*(I)$ that maximizes $W = E(\tilde{B}(D) - \tilde{C}(S) - \tilde{p}\tilde{I})$.

It is clear from this formulation that the problem of optimum tariff structure, as we have posed it, is yet another example of a wide class of optimum control problems with imperfect information which share a common structure. Other examples include the analysis of sharecropping with risk and incentive effects (for example, Stephen Cheung; Stiglitz, 1974); the choice of an optimum income tax structure (for example, Ray Fair; James Mirrlees, 1971, 1974, 1976; Eytan Sheshinski; Anthony Atkinson and Stiglitz; the analysis of insurance markets (for example, Kenneth Arrow; A. Michael Spence and Richard Zeckhauser; Mark Pauly; and Stiglitz, 1975c); the use of piece rates versus time rates in labor contracts (for example, John Pencavel, Stiglitz, 1975b); the determination of the optimum tariff structure for utilities (for example, Martin Weitzman, 1974a; Spence, 1975); the analysis of education as a screening device (for example, Spence, 1974; Stiglitz, 1975a); the question of using prices or quantities in planning (for example, Weitzman, 1974b; Mark Roberts and Spence). We have gone into the nature of the common structure of these seemingly diverse problems in Dasgupta and Stiglitz.

We should perhaps emphasize that we have restricted ourselves to the pure tariff and the pure quota cases not only because of ease of calculation. There is in addition the question of whether it is reasonable to imagine a government announcing and administering hopelessly complicated tariff schedules. Approximations to the optimum are therefore required, and it is this that is achieved by restricting the admissible set of tariff schedules to those that are simple in form. But still a fixed tariff may be unduly restrictive. One could well imagine, for example, that an appropriately chosen two-tier tariff structure would prove superior to the pure tariff rate t^*. Formally, such a schedule would read $t(I) = t_1$ for $I < \hat{I}$, and $t(I) = t_2$ for $I \geq \hat{I}$; normally $t_1 \neq t_2$ for an optimum; that is, an optimum two-tier tariff would be superior to a single tariff rate.

REFERENCES

Kenneth J. Arrow, *Essays on the Theory of Risk-Bearing,* Chicago 1971.

A. B. Atkinson and J. E. Stiglitz, "The Design of Tax Structure: Direct versus Indirect Taxation," *J. Publ. Econ.*, Aug. 1976, *5*, 283–301.

Jagdish N. Bhagwati, "On the Equivalence of Tariffs and Quotas," in Bhagwati et. al., eds., *Trade, Tariffs and Growth,* London 1969.

S. Cheung, "Transaction Costs, Risk Aversion, and the Choice of Contractual Arrangements," *J. Law Econ.*, Jan. 1969, *12*, 23–42.

P. Dasgupta and J. E. Stiglitz, "Incentive Schemes under Differential Information Structures: an Application to Trade Policy," IMSSS tech. rep. no. 172, Stanford Univ., July 1975.

R. C. Fair, "The Optimal Distribution of Income," *Quart J. Econ.*, Nov. 1971, *85*, 551–79.

J. A. Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev Econ. Stud.*, Apr. 1971, *38*, 175–203.

———, "Notes on Welfare Economics, Information and Uncertainty," in Michae Balch et al., eds., *Essays on the Theory o Economic Behaviour under Uncertainty* Amsterdam 1974.

———, "Optimal Tax Theory: A Synthesis," work. paper no. 176, Mass. Inst

Technology, May 1976.

M. **Pauly**, "Overinsurance and the Public Provision of Insurance," *Quart. J. Econ.*, Feb. 1974, *88*, 53–72.

J. **Pencavel**, "An Essay on the Economics of Work Effort and Wage Payments Systems," mimeo., Stanford Univ., Aug. 1974.

M. **Roberts and** A. M. **Spence**, "Effluent Charges and Licenses under Uncertainty," IMSSS, tech. rep. no. 146, Stanford Univ., July 1974.

E. **Sheshinski**, "The Optimal Linear Income Tax," *Rev. Econ. Stud.*, Oct. 1972, *39*, 297–302.

A. **Michael Spence**, *Market Signalling*, Cambridge 1974.

———, "Nonlinear Prices and Welfare," IMSSS, tech. rep. no. 158, Stanford Univ., Oct. 1975.

——— **and R. Zeckhauser**, "Insurance, Information and Individual Action," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 380–87.

J. **Stiglitz**, "Incentives and Risk Sharing in Sharecropping," *Rev. Econ. Stud.*, Apr. 1974, *41*, 219–56.

———, (1975a) "The Theory of Screening, Education and the Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

———, (1975b) "Incentives, Risk, and Information: Notes Towards a Theory of Hierarchy," *Bell. J. Econ.*, Autumn 1975, *6*, 552–79.

———, (1975c) "Monopoly and Imperfect Information: The Insurance Market," disc. paper, Stanford Univ., July 1975.

M. **Weitzman**, (1974a) "Is the Price System or Rationing more Effective in Meeting True Needs for a Deficit Commodity?," mimeo., Mass. Inst. Technology, Oct. 1974.

———, (1974b) "Prices vs. Quantities," *Rev Econ Stud.*, Oct. 1974, *41*, 477–92.

Land and Zoning in an Urban Economy: Further Results

By ELHANAN HELPMAN AND DAVID PINES*

In a recent paper, William Stull incorporates zoning considerations into an urban model. The specific issue discussed in his paper is the determination of the boundary between the manufacturing and residential zones. He assumes that the city is linear and its area exogenously subdivided into lots. The manufacturing zone includes the central lots, and the residential zones extend symmetrically from the manufacturing zone to the agricultural belt.

Zoning regulations which determine the boundary of the manufacturing zone are required to control the negative neighborhood externality effect of manufacturing on households. For simplicity, it is also assumed that these negative external effects increase with the proximity of residential location to the boundary of the manufacturing zone.[1]

Stull elaborates on the optimal zoning regulations which maximize the total revenue of a developer who owns all the land of an open city. Thus, the utility level of the population, which is assumed to be homogeneous, is exogenously given. The same is true with regard to the price of the product produced in the manufacturing zone.

When the developer extends the boundary of the manufacturing zone further away, the effects on his total revenue are the following: First, he loses the residential rent of the lots transferred from residential use to manufacturing.

Second, the negative neighborhood externality effect increases in the residential zone. Thus, given their utility level, households should be compensated by lower residential rent; otherwise they migrate to other cities. Consequently, total proceeds from residential rent decline.

Third, households located at the very boundary of the city, who already pay zero rent, migrate to other cities. Otherwise their utility level will be lower than what they can realize elsewhere. This reduces the supply of labor and increases the wage rate. Thus, the bid price for manufacturing lots, which is a decreasing function of the wage rate, declines too. Therefore, total revenue from the initial manufacturing zone decreases.

Fourth, the developer gains additional revenue which equals the manufacturing bid price of the lots transferred from residential to manufacturing uses.

The developer tends to extend the boundary of the manufacturing zone as long as the fourth (positive) effect exceeds the sum of the first three (negative) effects. Stull shows that in the developers' optimum, the producer's bid price for lots on the boundary of the manufacturing zone exceeds the households' bid price for lots on the boundary of the residential zone by an amount which is equal to the money value of the total negative neighborhood externality effect of a marginal extension of the manufacturing boundary.[2]

As we emphasized above, Stull's analysis is confined to the case of a developer who owns all the land and maximizes his profits. In the last section of the paper, he turns to the community case:

> "Suppose we now drop the assumption that all land in the city is owned and controlled by a private developer and

*Department of economics, Tel-Aviv University, and department of economics and Center for Urban and Regional Studies, Tel-Aviv University, respectively.

[1]In general, an external effect such as pollution depends on the level of concentration, which in turn depends on the level of activity of the polluting source and the distance of the damaged activity from the source.

[2]The third effect is not reflected in this cost-benefit calculation since it is of a secondary order of magnitude (see the discussion following equation (17))

instead assume decentralized ownership and community regulation. Is it reasonable under these circumstances to assume that the community will behave like the developer described above and choose the manufacturing-residential boundary point which maximizes aggregate land rent? This is in fact a complex question and one whose full exploration is beyond the scope of this paper." [p. 346]

In our earlier paper, we analyzed optimal zoning regulations for a system of open cities (optimal in the sense of utility maximizing). In this note, we show that Stull's conditions for maximum revenue of the developer coincide with the conditions of a social optimum. This implies that a social optimum is attained when each local authority imposes zoning regulations so as to maximize the land value in its jurisdiction.

I. A Restatement of Stull's Model

In what follows, we employ Stull's notation.

A

A household's utility function is[3]

$$(1)\quad u = u(x, q, t, t - t^*)$$

where

x = a composite good
q = land rented for housing
t = the distance from the consumer's residence to the urban center
t^* = the distance from the boundary of the industrial zone to the urban center

We use x as a numeraire good, so that all prices are in terms of this good.

Let $R(t)$ be the rent paid t miles from the center. Then, the compensated demand functions for the composite good and for land rented for housing are:[4]

$$x = x[1, R(t), t, t - t^*, u]$$
$$q = q[1, R(t), t, t - t^*, u]$$

Using these functions we define the minimum expenditure function:

$$(2)\quad \mu[R(t), t, t - t^*, u] \equiv x[1, R(t), t, t - t^*, u] + R(t)q[1, R(t), t, t - t^*, u]$$

$\mu(\ )$ gives the minimum expenditure in terms of x that a household needs in order to reach the utility level u while residing t miles from the center. From properties of the compensated demand functions, we have:

$$(3)\quad \mu_1 = q$$

$$(4)\quad \mu_3 = -u_4/\alpha$$

where α = the marginal utility of income

Given income y, an alternative utility level $\bar{u}$, and the boundary of the industrial area at t^*, a household which resides t miles from the center is willing to pay at most $m(t; y, t^*, \bar{u})$ units of x per unit of land in order to remain at its location. Hence, the $m(\ )$ function is implicitly defined by

$$(5)\quad \mu[m(t; y, t^*, \bar{u}), t, t - t^*, \bar{u}] + h(t) \equiv y$$

where $h(\ )$ = household transportation cost function. Equations (3), (4), and (5) imply

$$(6)\quad \partial m/\partial t^* = \mu_3/q = -u_4/(\alpha q)$$

$$(7)\quad \partial m/\partial y = 1/q$$

It is easy to see that Stull's multiplier λ equals $1/(\alpha q)$. Hence, (6) is consistent with Stull's equation (12).

Observe, however, that our formulation differs from Stull's in that we do not assume (as he did) that the residential zone is exogenously subdivided into equal lots. We let the household choose any desired lot at the going price. This seems to us a more reasonable specification, although our main result does not depend on it.

Now, if $q(t)$ is the amount of land rented by a single household t miles from the center, and $\bar{q}(t)$ is total usable land t miles from

[3]Stull considers x as a vector, but our simplification is immaterial.

[4]See Peter Diamond and Daniel McFadden for properties of compensated demand functions and the minimum expenditure function.

the center, then[5] $n(t) = \bar{q}(t)/q(t)$ is the number of households which reside t miles from the center and

$$(8)\quad N = \int_{t^*}^{t'} \bar{q}(t)/ q[1, m(t; y, t^*, \bar{u}), t, t - t^*, \bar{u}]\, dt$$

is total population, which equals labor supply (t' is the city's boundary).

B

Let the production function be

$$(9)\qquad Z = F[L(t), \bar{q}(t)]$$

It takes $k(t)$ units of the composite good to ship one unit of the composite good from its production point t miles from the center to the center. Given the wage rate w, the maximum rent per unit land that a firm occupying a ring at t is willing to pay (i.e., its bid prices) is:[6,7]

$$(10)\quad M(t, w) \equiv \max_L \{[1 - k(t)]F[L, \bar{q}(t)] - wL\}/\bar{q}(t)$$

Let $L(t, w)$ be the solution to the right-hand side of (10). Then we have

$$(11)\quad [1 - k(t)]F_1[L(t, w), \bar{q}(t)] \equiv w$$

$$(12)\quad M(t, w) \equiv \{[1 - k(t)]F[L(t, w), \bar{q}(t)] - wL(t, w)\}/\bar{q}(t)$$

Equations (11) and (12) imply[8]

$$(13)\quad \partial M/\partial w = -L(t, w)/\bar{q}(t)$$

$$(14)\quad \partial M/\partial t = -k'(t)F[L(t), \bar{q}(t)]/\bar{q}(t)$$

[5]In a linear city $\bar{q}(t)$ is fixed for every t. Stull assumed a linear city, but we allow for more general forms.

[6]Stull identifies income (y) with the wage rate (w). We argue in the next section that—from general equilibrium considerations—this identification is incorrect.

[7]See Stull's equation (21).

[8]See Stull's equation (24). Equation (14) is obtained by either assuming constant returns to scale and/or by assuming a linear city; i.e., $\bar{q}(t)$ = constant. However, (14) is not required for the derivation of our main result; it is presented here only for comparison with Stull's equation (24).

C

A household's income is composed of wages and nonwage income. Hence[9]

$$(15)\qquad y = w + z$$

where $\qquad z$ = nonwage income

The equilibrium wage rate is determined so as to clear the labor market. The labor market-clearing condition is (see (8) and (11))

$$(16)\qquad N = \int_0^{t^*} L(t, w)\, dt$$

N depends also on w, and (16), (15), and (8) provide an implicit solution for w. This solution depends generally on t^* and t', but this relationship is not needed to be known for our purpose. Let it be denoted by $w(\ )$.

The developer takes the nonwage income z, and the utility level $\bar{u}$, as given. He chooses t^* and t' so as to maximize

$$\int_0^{t^*} M[t, w(\ )]\bar{q}(t)\, dt + \int_{t^*}^{t'} m[t; w(\ ) + z, t^*, \bar{u}]\bar{q}(t)\, dt$$

The first-order condition with respect to t^* is

$$(17)\qquad M(t^*)\bar{q}(t^*) = m(t^*)\bar{q}(t^*) - \int_{t^*}^{t'} \partial m/\partial t^* \bar{q}(t)\, dt - \left\{\int_0^{t^*} \partial M/\partial w \bar{q}(t)\, dt + \int_{t^*}^{t'} \partial m/\partial y \bar{q}(t)\, dt\right\} \frac{dw}{dt^*}$$

Equations (7), (8), (13), and (16) imply th the last term on the right-hand side of (1 disappears. Hence

$$(18)\quad M(t^*)\bar{q}(t^*) = m(t^*)\bar{q}(t^*) - \int_{t^*}^{t'} \partial m/\partial t^* \bar{q}(t)\, dt$$

For a linear city, (18) reduces to Stull equation (36).

[9]Stull assumes $z = 0$. See also fn. 6.

II. Optimal Zoning: The Community Case

Now consider a system of I cities which has to house an urban population of a given size. Each city has a basic structure as the city that was described in the previous section, but cities may differ in utility functions, production functions, household transportation functions, commodity transportation functions, usuable land functions, and alternative land costs.[10] We denote by a superscript i, $i = 1, 2, \dots, I$, the relevant variables and functions for city i.

The urban population is assumed to be homogeneous in preference, initial holdings, and productivity. In particular, each household has an equal share in each land dealing corporation. The land dealing corporations, which behave competitively, rent land from agriculture (possible at zero cost, as in Stull) and rent it out to urban users so as to maximize profits. All other economic agents are competitors too.

Households choose their location, both within and among cities, so as to maximize utility. This implies in equilibrium equalization of the utility level across cities as well as within cities.

Given this structure, what are the zoning regulations which maximize the common utility level? An answer to this question is provided in our earlier paper. It is shown there that optimal zoning satisfies

$$(19)\quad [1 - k^i(t^{*i})]\{F^i[L^i(t^{*i}, w^i), \bar{q}^i(t^{*i})] - L^i(t^{*i}, w^i)F_1^i[L_i(t^{*i}, w^i), \bar{q}^i(t^{*i})]\} = R^i(t^{*i})\bar{q}^i(t^{*i}) - \int_{t^{*i}}^{t^{\prime i}} \mu_3^i[R^i(t), t, t - t^{*i}, \bar{u}]n^i(t)\,dt \qquad i = 1, 2, \dots, I$$

where $R^i(t)$ is the urban rent gradient which satisfies:

$$(20)\quad \mu^i[R^i(t), t, t - t^{*i}, \bar{u}] + h^i(t) = w^i + z, \quad i = 1, 2, \dots, I, t \in [t^{*i}, t^{\prime i}]$$
$$n^i(t) = \bar{q}^i(t)/q^i[1, R^i(t), t, t - t^{*i}, \bar{u}]$$

(the number of households that reside t miles from the center in city i) and $\bar{u}$ is the common maximal utility level.

Equation (19) has a very simple cost-benefit interpretation. Suppose we consider an extention of city i's industrial area boundary by a marginal unit. Then, the left-hand side of (19) represents the net addition to the city's output that will result from this action.[11] This is the benefit. The right-hand side represents costs. The first term represents the alternative value of land that is transferred from housing and/or agriculture to industry. The second term represents the value of the increased neighborhood externality effect that results from the fact that each household now has to live closer to the industrial area. It equals the city household's aggregate willingness to pay for the prevention of the marginal change in the industrial area's boundary.

Equation (20) implies that the rent gradient $m(t; w^i + z, t^{*i}, \bar{u})$ equals $R^i(t)$ (see (5)), provided that the optimal values of $w^i + z$, t^{*i} and $\bar{u}$ are used in $m(\ )$. Observe also that the wage rate may differ

[10]Stull assumes that alternative land costs for urban uses equal zero.

[11]To see this, consider the following indirect production function (we omit the superscript i):

$$f(L, t^*) \equiv \left\{ \max_{L(t)} \int_0^{t^*} [1 - k(t)]F[L(t), \bar{q}(t)]\,dt; \quad \text{such that} \quad \int_0^{t^*} L(t)\,dt = L \right\}$$

Then, it can be easily verified (by either direct calculations or by an application of the Envelope Theorem) that:

$$\partial f(L, t^*)/\partial t^* = [1 - k(t^*)]\{F[\bar{L}(t^*), \bar{q}(t^*)] - \bar{L}(t^*)F_1[\bar{L}(t^*), \bar{q}(t^*)]\}$$

where $\bar{L}(t)$ = the output-maximizing value of $L(t)$ and $[1 - k(t)]F_1[\bar{L}(t), \bar{q}(t) = \lambda$ = the constraint's multiplier. This result can be explained as follows. Suppose we add an additional ring of width dt to the manufacturing zone. Output (net of shipment costs) increases by $[1 - k(t^*)]F[\bar{L}(t^*), \bar{q}(t^*)]dt$, if we reallocate labor in an efficient way. But this means that we take $\bar{L}(t^*)dt$ workers from other rings and put them to work in the new ring. We lose, therefore, the marginal output of these workers, which equals $\bar{L}(t^*)\ \lambda\ dt$. Hence, the net increase in output is $[1 - k(t^*)]F[\bar{L}(t^*), \bar{q}(t^*)]dt - \bar{L}(t^*)\lambda dt = [1 - k(t^*)]\{F[\bar{L}(t^*), \bar{q}(t^*)] - \bar{L}(t^*)F_1[\bar{L}(t^*), \bar{q}(t^*)]\}dt$. More generally, given an aggregate production function $f(L, t^*)$, the left-hand side of (19) should be $\partial f(L, t^*)/\partial t^*$.

among cities, but the nonwage income should be the same in each city. The nonwage income is equal to total rent plus net pure profits divided by the total number of urban population (see the authors).

Now, from (6) we obtain $\bar{q}(t)\,\partial m(t)/\partial t^* = \mu_3(t)n(t)$. Hence, (18) and (19) are identical if

$$(21)\quad M(t,w)\bar{q}(t) \equiv [1-k(t)]\{F[L(t,w),\bar{q}(t)] - L(t,w)F_1[L(t,w),\bar{q}(t)]\}$$

But from (11) and (12) it is clear that (21) is satisfied. Hence, land value maximization is efficient from the social point of view if the developer (or for this purpose the local authority) uses the bid prices for land that were specified by Stull; provided, of course, that the nonwage income is properly accounted for.

This result is in line with recent findings of Pines and Yoram Weiss. They showed that in an open city without production, the marginal benefit of a policy measure is fully reflected in the city's total land value increase, where competitive prices are used to calculate land values. This implies that land value maximization is the appropriate policy criterion in this setting, provided that the policy itself is costless (as in the case of zoning).

The Pines-Weiss result can be generalized as follows. It can be shown that in a more general setting—where production, for example, also takes place—the marginal benefit of a policy measure is fully reflected in the city's total pure profits increase, where pure profits consist of pure rents (or profits of land dealing corporations) and profits from production. This means that for overall efficiency, an open city should choose its costless policy measure so as to maximize the city's pure profits, and that it should choose its costly policy measures (like the provision of public goods) so as to equate the marginal increase in pure profits with the marginal cost of the policy measure. This is indeed done by Stull's developer. Although the developer maximizes land values, his procedure is equivalent to total profit maximization, since the producer's bid price for land consists of the marginal product of land plus pure profits per unit land. Hence, the developer's total land value comprises, in fact, pure rents plus pure profits from production, and this has indeed to be maximized in order to assure efficiency.

Competition fails to generate an efficient allocation because it fails to internalize the neighborhood externality effect, which can also be looked upon as a negative externality of producers on land dealers. In this case total profits are not maximized.

Finally, it is worth noting that there is an assymmetry in Stull's developer's treatment of households and firms. When calculating bid prices the developer takes away from firms the producer's surplus while he does not do it with respect to households. Were the developer calculating the households' bid price so as to include the consumer's surplus, the resulting zoning would not be efficient.

REFERENCES

P. A. Diamond and D. L. McFadden, "Some Uses of the Expenditure Function in Public Finance," *J. Publ. Econ.*, Feb. 1974, *3*, 3–21.

E. Helpman and D. Pines, "Optimal Zoning and Corrective Taxation in a System of Open Cities," work. paper no. 90, Foerder Inst. Econ. Res., Tel-Aviv Univ., rev. June 1976.

D. Pines and Y. Weiss, "Land Improvement Projects and Land Values," *J. Urban Econ.*, Jan. 1976, *3*, 1–13.

W. J. Stull, "Land Zoning in an Urban Economy," *Amer. Econ. Rev.*, June 1974, *64*, 337–47.

Pareto-Desirable Redistribution in Kind: An Impossibility Theorem

By GEOFFREY BRENNAN AND CLIFF WALSH*

In recent years there has been a considerable, and growing, interest in the possibility of applying the familiar Paretian welfare framework to questions of income redistribution. In contrast to the more traditional treatments of redistribution in which "donor-taxpayers" are taken to be characteristically *unwilling* participants in the redistributive process,[1] the central observation in this new approach is that redistribution may for a variety of reasons yield benefits to donors as well as to recipients. Thus, donor-taxpayers may have philanthropic inclinations towards the poor; or they may contribute to transfer programs as a way of insuring themselves against future income loss; or they may seek by redistribution to avoid the social and political unrest associated with large inequities in the income distribution.[2]

One of the implications of the Paretian perspective on distributional questions, with its focus on donor-taxpayer preferences, has been to raise questions about the form which interpersonal transfers should take. In particular, the traditional view that redistribution should be effected in a lump sum manner (or the feasible best approximation to it), so as to avoid distorting the choices of recipients, seems to have given way to the view that in many cases redistribution in kind (of the per unit subsidy type) may be required because donor-taxpayers may not be indifferent as to how recipients spend the transfers they receive.[3] In fact, it has become more or less customary in the literature to distinguish quite sharply between generalized philanthropy on the one hand and specific commodity interdependence on the other as alternative types of justification for redistribution; indeed, in large measure the two approaches have developed independently.[4] Thus, as the literature now stands, the Pareto-desirable redistribution possibility appears to be firmly and equally supported by the twin pillars of general and specific forms of interdependence.

Our objective in this brief note is to demolish one of these pillars. Specifically, we argue that, for all intents and purposes, there is no such thing as Pareto-desirable redistribution in kind—that whether the position taken on income redistribution is Paretian or not, the appropriate form of redistribution is lump sum.

At the outset it should perhaps be emphasized that when we refer to redistributive activity as being "Pareto desirable" (or more loosely, "Pareto optimal"), we mean that redistribution is *required* by the Pareto criterion. We would distinguish sharply between redistribution which is Pareto desirable in this sense, and that which is merely *consistent with* the Pareto criterion.

*Senior lecturer and lecturer in public finance, Australian National University, respectively. The final version was prepared while Brennan was a visiting research associate at the Center for Study of Public Choice, Virginia Polytechnic Institute and State University. We are grateful to an anonymous referee for helpful comments.

[1]The redistribution involved might be viewed as either arising out of the manipulations of some "ethical observer," or simply emerging as the result of a zero-sum game between income classes in the political arena: but in all cases donor-taxpayers are taken to be made worse off by the transfer process.

[2]Analysis and discussion of the basic philanthropy model can be found in Harold Hochman and James Rodgers (1969), George von Furstenberg and Dennis Mueller, and in the authors. Other cases are considered in Brennan (1973, 1975).

[3]Throughout this paper "in-kind distribution" is defined as redistribution which induces a substitution effect. As is widely acknowledged, noncash transfers do not always satisfy this requirement.

[4]References to the former approach are to be found in fn. 2. For the latter approach the relevant works include James Buchanan (1959, 1968), Edgar Olsen (1969, 1971), Rodgers (1973), George Daly and Fred Giertz (1972, 1976), and George Peterson.

To clarify this distinction, we should note that it is possible to make interpersonal transfers in association with *any* Pareto-desirable move and still leave everyone better off: what would be achieved would be a redistribution of the "gains from trade" involved in that move. But the *redistribution* in this case could not be described as Pareto desirable because it cannot in any way be regarded as necessary for the achievement of a Pareto optimum: it is simply an optional extra.

If it can be agreed that what we are concerned about are situations in which efficiency considerations *require* redistribution, then we can characterize *all* Pareto-desirable redistribution situations as being associated with an upward-sloping utility possibilities frontier over some range.[5] Such an upward-sloping segment implies that, over that range, in order to make one party (the donor) better off it is *necessary* to make the other party (recipient) better off as well.

I

With this as background, consider first the situation that appears to have been typically taken as providing the basis for in-kind redistribution—that in which donor-taxpayers have preferences relating to how "recipients" spend their incomes, including any transfers received. This is usually taken to imply that "donors" have a utility function of the form:

$$(1) \quad U_D = U_D(X_D^1, X_D^2, \ldots, X_D^N; X_R^1)$$

where X_D^i are goods consumed by D and X_R^i are goods consumed by R.

Each recipient R is taken to have a utility function which includes only goods consumed by himself, that is,

$$(2) \quad U_R = U_R(X_R^1, \ldots, X_R^N)$$

Such a situation is of course nothing more nor less than a standard externality of the nonreciprocal type (as analyzed for example by Buchanan and W. C. Stubblebine). As is well known, optimality requires that R consume X^1 not at the point where

$$(3) \quad MRS^R_{X^1X^N} = MRT_{X^1X^N}$$

but rather, where

$$(4) \quad MRS^R_{X^1X^N} = MRT_{X^1X^N} - MRS^D_{X^1_RX^N}$$

It is also well-known that the relevant substitution effect needed to achieve (4) in R's consumption can be effected via a subsidy on X_R^1, paid by D: hence, the redistribution possibility. But equally, (4) can in general be achieved via a *tax* on R's consumption of all other goods ($X_R^2, \ldots X_R^N$), a fact which is also widely recognized in the literature.[6,7] In using the latter possibility, it may of course be necessary to ensure—in obedience to the Pareto criterion—that R is made no worse off. But the crucial point is that there is nothing inherent in this situation which ensures that R must be made *better* off—the Pareto criterion does not insist that redistribution in R's favor should take place, and any "redistribution" which happens to occur cannot, therefore, be described as *Pareto desirable.*[8]

To put the point a slightly different way,

[5]This point is already widely accepted in the literature. See, for example, Hochman and Rodgers (1969, p. 999) or Brennan (1975).

[6]Edward Mishan and Mark Pauly are perhaps particularly relevant in this context, though neither see the full implications of the point.

[7]Of course, it is conceivable that the externality would still remain Pareto relevant even when R spent *all* his income on X^1, in which case some redistribution to R might be said to be implied by (4). But such corner solutions do not seem to be what the literature has had in mind, and in any case one can have doubts about their practical relevance.

[8]There is actually a problem in determining how redistribution ought to be defined in this context. Suppose that the externality is internalized by a subsidy paid by the "affected" to the "affecting" parties. It is obvious that the payment of this subsidy makes the affecting party (the recipient) better off. But it is not entirely obvious that this process can be appropriately described as redistribution. For when two individuals enter the market place to trade completely private goods, both (including specifically the poorer) are made better off—but do we (or should we?) refer to this trade as involving redistribution? It would certainly seem to be contrary to standard economic terminology to do so. Yet one of the things that the postwar externality literature has emphasized is that the process of internalizing standard externalities is exactly analogous to a trade of private goods (or of private property rights in the relevant asset). Thus,

to use a Pigovian subsidy to achieve optimality (equation (4)) in this situation involves the creation of both an income effect and a substitution effect. Only the substitution effect is required to achieve optimality, but only the income effect involves redistribution. Since the income effect is *not* required by the Pareto criterion, the redistribution is *not* Pareto desirable.[9]

II

Of course, it is possible to reformulate the situation so as to ensure that the income effect *is* necessary to achieve optimality. This can be done by incorporating into the donor-taxpayer's utility function a *general* feeling of concern for R as well as a preference about the way in which R spends his income. Accordingly, we have:

$$(5) \quad U_D = U_D(X_D^1, \ldots, X_D^N; X_R^1, U_R)$$

with R's utility function as set out in equation (2). Given that both the specific and the general utility interdependencies are Pareto relevant, this formulation is sufficient to ensure that both a substitution *and* an income effect are required for optimality. Under certain circumstances, the substitution effect and the income effect required for optimality would both be such as to be exactly satisfied by some per unit price subsidy arrangement. In others a rather larger (or rather smaller) income effect will be required than can be conveniently achieved through a simple reduction of the price of X_1 to R. In all cases however some part of the redistribution involved will be *required* by the Pareto criterion rather than simply consistent with it; and this redistribution *can* be appropriately described as Pareto desirable (or Pareto optimal or just "Paretian" as whim decrees). Such Pareto-desirable redistribution is of course the redistribution necessary to internalize the generalized concern which D has for R, indicated by the presence of U_R in D's utility function. Any redistribution beyond this amount is either attributable to non-Paretian judgments about the distribution of gains from trade, or is a happy (or unhappy) accident—either way, it is *not* Pareto desirable.

But if it is the presence of U_R in D's utility function, and that only, which gives rise to Pareto-desirable redistribution, then it is the internalization of that interdependence which determines the form which Pareto-desirable redistribution must take. And of course *that* interdependence can only be internalized via *lump sum transfers.*

None of this is to deny the prevalence of standard externalities, or the need for changes in relative prices to internalize them. What is being denied is that any *income redistribution* involved in these relative price changes can be described as Pareto desirable. Pareto-desirable redistribution can arise only if *income* effects are required for optimality: and, not surprisingly, income effects are achieved by income transfers.

Thus, whether the perspective taken on income distribution is Paretian or not, it remains true that:

(i) All specific good externalities of the standard type require substitution effects for optimality to be achieved; and

(ii) Any redistribution *required* should be implemented in lump sum form.

III

In this brief note we have attempted to indicate that the standard formulation of Pareto-desirable redistribution in kind in the literature is seriously deficient. We have also sought to offer an alternative reformulation, which seems to us to capture the essential spirit of what the literature has had in mind. This reformulation involves donor concern for *both* the level of recipient con-

paying an individual to collect his own garbage, paint his house, feed his children or whatever, could be looked on in the same way as paying him for services provided specifically in the labor market. It is a neat semantic conundrum precisely how wages are to be distinguished from these other payments. Or are wages redistribution as well?

[9]Put yet another way, there is nothing in the standard externality situation which will ensure an upward-sloping segment in the utility-possibilities frontier over the relevant range.

sumption of some particular good *and* for the recipient's well-being more generally. Unfortunately, under such a reformulation any *Pareto-desirable* redistribution must be effected in cash—the possibility of Pareto-desirable redistribution of an in-kind variety is obliterated!

We should perhaps emphasize that we are not seeking to imply by this that price subsidies (or redistribution in kind) cannot serve to make everyone better off, and hence that they cannot satisfy the Pareto test. What we *are* seeking to make clear is that such price subsidies have in general nothing to do with *Pareto-desirable redistribution*, and, in fact, that Pareto-desirable redistribution *in kind* is conceptually impossible.

REFERENCES

H. G. Brennan, "Pareto-Desirable Redistribution: The Non-Altruistic Dimension," *Publ. Choice*, Spring 1973, *14*, 43–67.

———, "'Pareto-Optimal Redistribution': A Perspective," *Finanzarchiv*, 1975, *33*, 237–71.

——— **and C. Walsh,** "Pareto-Optimal Redistribution Revisited," *Publ. Finance Quart.*, Apr. 1973, *1*, 147–68.

J. M. Buchanan, "Positive Economics, Welfare Economics, and Political Economy," *J. Law Econ.*, Oct. 1959, *2*, 124–38.

———, "What Kind of Redistribution Do We Want?," *Economica*, May 1968, *35*, 185–90.

——— **and W. C. Stubblebine,** "Externality," *Economica*, Nov. 1962, *29*, 371–84.

G. Daly and F. Giertz, "Welfare Economics and Welfare Reform," *Amer. Econ. Rev.*, Mar. 1972, *62*, 131–38.

——— **and** ———, "Transfers and Pareto Optimality," *J. Publ. Econ.*, Jan./Feb. 1976, *5*, 179–182.

H. M. Hochman and J. D. Rodgers, "Pareto Optimal Redistribution," *Amer. Econ. Rev.*, Sept. 1969, *59*, 542–57.

——— **and** ———, "Pareto Optimal Redistribution: Reply," *Amer. Econ. Rev.*, Dec. 1970, *60*, 997–1004.

E. J. Mishan, "Redistribution in Money and Kind: Some Notes," *Economica*, May 1968, *35*, 191–93.

E. O. Olsen, "A Normative Theory of Transfers," *Publ. Choice*, Spring 1969, *6*, 30–57.

———, "Some Theorems in the Theory of Efficient Transfers," *J. Polit. Econ.*, Jan./Feb. 1971, *79*, 166–76.

M. V. Pauly, "Efficiency in the Provision of Consumption Subsidies," *Kyklos*, Mar. 1970, *23*, 33–57.

G. E. Peterson, "Welfare, Workfare and Pareto Optimality," *Publ. Finance Quart.*, July 1973, *1*, 323–38.

J. D. Rodgers, "Distributional Externalities and the Optimal Form of Income Transfers," *Publ. Finance Quart.*, July 1973, *1*, 266–99.

P. A. Samuelson, "A Diagrammatic Exposition of a Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1955, *37*, 350–56.

G. M. von Furstenberg and D. C. Mueller, "The Pareto Optimal Approach to Redistribution: A Fiscal Application," *Amer. Econ. Rev.*, Sept. 1971, *61*, 628–37.

Comparing Utility Functions in Efficiency Terms or, What Kind of Utility Functions Do We Want?

By BURTON A. WEISBROD*

This note considers a hypothetical experiment designed to give meaning to the concept of one "type" of utility function being "preferred" to another. In the course of presenting the outlines of such a conceptual experiment, the terms type and preferred will be defined. The conclusion—albeit at a quite abstract level—is that while there exists a class of preferences that are indeed noncomparable, as contemporary economic theory holds, there also exists a class of preferences that can be compared with each other. With respect to the latter class we find that there is a significant sense in which one set of preferences (and the expected consumption bundle associated with it) can be said to be *preferred* to another set of preferences (and the expected consumption bundle associated with it). That being the case, it becomes meaningful to talk about the efficiency or inefficiency of allocating resources to shaping preferences.[1]

I. A Hypothetical Experiment

1. Each participant, each prospective member of society, is asked to vote—that is, to express a preference—on whether he would prefer to live in a society with given resource endowments in which everyone, including himself, had a type I or, alternatively, a type II utility function (to be defined below).

2. Each participant is asked to vote from behind what John Rawls has termed a "veil of ignorance" such that he does not know the "initial conditions" with which he would enter the society. At this "constitutional" level he therefore has the average expectation of all members of the society as to how "well off" he would be in material and nonmaterial terms in each "state" of society; a state is defined as the expected consumption bundle resulting from an economic system, with given resources and technology, in which people had a particular type of utility function.[2]

3. The participant is asked the following two conditional questions (or, in general, *n* questions, one for each type of preference function and accompanying state of society, from which the social choice is to be made):

(a) If you were in the type of society in which everyone, including you, behaved according to a type I utility function, and had the consumption bundle associated with that society, would you prefer to see a change to another type of society, in which everyone behaved according to a type II utility function and had the associated consumption bundle?

(b) Conversely, if you were in a type II society, would you prefer to see a change to type I?

*Professor of economics and fellow, Institute for Research on Poverty, University of Wisconsin-Madison. I want to acknowledge the helpful comments of Richard Dusansky, A. James Lee, Mancur Olson, Efraim Sadka, and Eugene Smolensky on an earlier draft. I want to give special thanks to Arthur Snow not only for his valuable comments on previous drafts but also for the many productive conversations we have had in the process of developing this paper. He has contributed significantly to it.

[1]In a recent paper Roland McKean characterizes contemporary thinking in economics on the noncomparability of utility functions: "Economics provides no analytic framework for saying anything about the worth of alternative preferences . . . ." (p. 640).

For some recent analyses of preferences as variables see Carl Weizsäcker, M. McManus, and Amartya Sen.

[2]When I say that the participant has the "average expectation" of all members of the society I do not intend to imply that Rawls views the decision process in such probabilistic terms. (I am indebted to James Fishkin for suggesting this clarification.)

4. If and only if the answers to both these conditional questions (3a and 3b) were consistent in expressing a preference for either type I or II preferences would we conclude that one type of utility function (and the associated consumption bundle) is preferred by that person. Otherwise the preferences would be deemed noncomparable. In other words, if a person prefers to be in a society in which everyone has one particular type of utility function (say type II) rather than another (say type I)—prefers it in the sense that he prefers the consumption bundle he would expect to receive in society II to the bundle he would expect to receive in society I, and he has this preference ordering no matter which type of utility function he uses as the basis for his evaluation—then he would be said to prefer a type II utility function (or a type II society) to a type I. This will be referred to below as the rule 4 definition.

Now consider two cases in which different pairs of responses were given to questions 3(a) and 3(b):

CASE A: The only difference between states I and II (i.e., between the types of utility functions from which a choice can be made) is that in I, everyone prefers eating meat to eating fish, while in state II, everyone prefers fish to meat. A person who envisages himself being in state I, having a preference for meat, is asked (question 3, above) whether he would prefer to be in (i.e., would vote for) another society in which everything was the same except that he and everyone else preferred fish. There is no reason to believe, given the structure of this example, that a change would be preferred. A person who is a meat lover would have no reason to want to become a fish lover, and vice versa, when, as proposed here, his decision is based only on an evaluation of alternative consumption bundles that are feasible in the two types of societies—for these bundles are the same. The two types of utility functions, differing only in marginal rates of substitution among privately consumed goods, are *not comparable* in that individuals, from behind a veil of ignorance, would not prefer to change types of utility functions, no matter which they might initially have. Thus, in case I we would have the following answers to questions 3(a) and 3(b):

If in state I, then state I (and its expected consumption bundle) *is preferred to state II; and if in state II, then II* (and its expected consumption bundle) *is preferred to I.* According to rule 4, above, since the typical individual's answers are not consistent, we would say that the two alternative preferences are noncomparable.

This is the standard textbook case. The individual's consumption possibilities are independent of the type of preference function he and others have. There is no basis for proposing a voting rule for choosing among utility functions—at least no basis that relies on the weak ethical criteria that the new welfare economics has accepted.

CASE B: The difference between type I and type II utility functions in Case B (referred to here as I′ and II′ to distinguish them from the Case A situation) is that I′ is what we term a "private" type utility function—each person i has a utility function that depends only on his own consumption, narrowly defined—$u_i = u_i(q_{1i}, q_{2i}, \ldots, q_{ni})$—and II′ is what we term the "internalized" type utility function $v_i = v_i(q_{1i}, q_{2i}, \ldots q_{ni}; z)$—where z is an argument that relates to the real external effects (positive or negative) of person i's activities. For example, z might reflect the desire to be honest in dealing with other people, or a preference for not imposing littering costs on other people.[3]

A person choosing between societies with type I′ or II′ utility functions would recognize, assuming he were well informed, that in a society of persons with type I′ utility functions, some of the society's scarce resources would have to be devoted to coping with the external diseconomies that narrow

[3]Conversely, z might also reflect a desire to do harm to others, but a utility function reflecting such a preference would not be one that internalized external effects.

self-interest together with nonzero enforcement and transactions costs will create. These diseconomies include, for example, pollution, litter, and failure of individuals to disclose information that would be valuable to others but somewhat harmful to the discloser.

By contrast, depending on the precise form of the type II′ utility function, the resource costs of coping with (Pareto-relevant) external diseconomies would tend to be smaller, and as a result consumption possibilities would tend to be greater than in a type I′ society having the same initial endowments. *Consumption possibilities are not independent of preferences.* Thus, the following result is possible (but not necessary) as a statement of the answers to questions 3(a) and 3(b):

If in state I′, then state II′ is preferred to I′; and

if in state II′, then state II′ is preferred to I′.

This says that at the constitutional or veil of ignorance level, it is *possible* that each person would feel that (a) if he were in a society of people with narrow-type utility functions, he would prefer that all people, including himself, had internalized-type utility functions, for they would then engage in less behavior that generates real external diseconomies, with the result that his expected consumption bundle would be "larger"; and (b) if he were in a society in which people had the internalized-type utility functions, he would prefer not to change to the type of society and the associated consumption bundle, in which private-type utility functions prevailed. Then according to rule 4 above, preferences of type II would be said to be preferred to those of type I.

II. Discussion

Nothing in the preceding section implies that a type II′ utility function and its accompanying state of society is necessarily preferred to its type I′ counterpart. All that is being said is that because of the greater external costs in the type I′ world, and the real costs of developing arrangements to internalize those costs, the type II′ utility functions *might* be preferred in the sense defined.[4]

The rule 4 definition may be restated: one type of utility function is said to be preferred (and, hence, comparable) to another if and only if the "base reversal" test is passed—that is, whichever utility function is utilized for evaluating the consumption bundles that would be expected in each type of society, the consumption bundle expected from one particular type of society is preferred. We say that one type of utility function is preferred to another if, from behind a veil of ignorance as to which type of utility functions were to characterize the society, a typical person would choose as follows: (a) if he and other members of society were to be "dealt" utility functions of type II′, he would vote to retain that type of utility function; *and* (b) if they were "dealt" utility functions of type I′, the typical person would vote to switch to type II′ functions.

The evaluation of a person's expected welfare under type I′ or II′ conditions is analogous to the evaluation of whether a "price level" is or is not "higher" at time 2 than at time 1; if and only if the price level is higher at time 2 *whether period 1 or period 2 weights are used* do we say that prices are unambiguously higher. Similarly, one type of utility function, and the expected consumption bundle it generates, may be said to be preferred to another, and the expected consumption bundle it generates, if and only if (a) the two expected consumption bundles are different and (b) the same consumption bundle is preferred no matter which utility function is used to evaluate the two bundles.

The distinction between the two types of utility functions that were compared in

[4]Just as resource costs of coping with external costs would be smaller in a society with type II′ utility functions, so, too, resource costs would be smaller if individuals with type I′ utility functions were more stoic, and thus were relatively unaffected by the behavior of others. (This point was made to me by Susan Rose-Ackerman.)

Cases A and B, above, is that in Case B, one of the utility functions had the effect of saving resources by internalizing costs that would not have been internalized without the utilization of scarce resources. In Case A, the two utility functions that were compared did not differ in their responses to external effects; they differed only in the resource allocations that they implied as between alternative consumption goods but did not differ in the degree to which external effects were internalized.

It is not true that even in the Case B choices, the type II′ utility functions would necessarily be preferred. People might prefer the type I′ utility function despite the result that the society it would imply would either have to bear the full cost of the external effects, or else would have to use resources to deal with or prevent the privately efficient but socially inefficient actions that the narrow self-interest model (type I′) would produce. (Note, however, that everyone would not necessarily be better off in a type II′ state even if the mathematically expected outcome is greater for that state.)[5] The important conclusion is not that one type of utility function *would* be preferred to another, but that it *could* be—that is, we can meaningfully talk about, analyze, and indeed compare alternative utility functions and the states of the world they imply. There is thus an important sense in which utility functions can be compared.

If utility functions can sometimes be compared, and one type can sometimes be said to be preferred to another, then the means for affecting the kinds of utility functions that people have becomes, at least in part, a matter of resource allocation: how much of its resources would it be efficient (in the Pareto sense) for society to allocate for the purpose of *shaping* utility functions if the alternative is devoting resources to pricing, taxing, subsidizing, and enforcing legal arrangements to deal with allocational inefficiencies resulting from noninternalized externalities?[6]

The quite abstract perspective of this paper can be given somewhat more concreteness by considering the case of young children, whom we may consider as persons whose adult utility functions are yet to be shaped. Society can and does make decisions regarding the development of these utility functions. In effect we might assume that a child is born with a "blank" utility function, or that society acts as if that were the case. The education and the religious systems, for example, can be and are used to shape preferences. The argument of this paper is that, conceptually, such preference formation can, within limits, be analyzed within a conventional allocative efficiency framework.

III. Conclusion

The customary proposition that one type of utility function cannot be compared to another within an economic efficiency framework is correct in general. This note has suggested, however, that despite the general noncomparability of utility functions, some can be compared and found to be preferred to others. In particular, individuals' utility functions that are of the "in-

[5] If some persons adopted internalized utility functions while others did not, the former group might or might not actually be better off unless compensation were paid, but such redistributional considerations are separable from efficiency considerations. With respect to efficiency, the argument here is analogous to the conventional "second best" analysis, just as allocative efficiency is not necessarily enhanced when one sector of an imperfectly competitive economy becomes more competitive, so it is not *necessarily* efficient when one subset of the population adopts internalized utility functions.

[6] Kenneth Arrow has implicitly recognized this in the context of the development of "ethical codes." He has argued that, "... ethical codes can contribute to economic efficiency" (p. 317). He did not point out explicitly, however, that the acceptance of ethical codes involves shaping or altering utility functions, and thus, to say that ethical codes can contribute to efficiency is equivalent to saying that some utility functions (those that reflect acceptance of certain ethical codes) are superior in efficiency terms to others. It is also equivalent to saying that it can be efficient to use resources for gaining acceptance of ethical codes and, thus, to shape utility functions. Arrow addressed his remarks to situations involving transmission of information, but his argument actually applies more broadly to all Pareto-relevant real externality situations.

ternalized" type, internalizing what would otherwise be external effects and thus saving resources, may be comparable with private-type utility functions, which disregard external effects and thus imply smaller expected consumption bundles. In short, some utility functions can be said to be preferred to others in the following sense: the expected consumption bundle in a society of people who have one type of utility function may be preferred to the expected consumption bundle in a society of people who have another type of utility function—using a base-reversal definition of what is preferred. Insofar as such preferences exist, it can be efficient to devote resources to shaping utility functions. Such a use of resources is an alternative to regulatory and tax subsidy mechanisms for internalizing what would otherwise be Pareto-relevant external effects.[7]

Recognition of the possibilities of comparing utility functions in conventional economic efficiency terms, and of shaping or reshaping utility functions as a potentially efficient alternative to taxes, subsidies, or regulation, may permit expansion of the domain of policy statements that economists can make within our familiar Pareto-welfare economic framework. It remains for additional research to explore the operational means for determining preferences and the costs and benefits of changing them. The main point is that the injunction to avoid trying to compare preferences on grounds that it is "unscientific" has been overstated. Using a quite weak ethical criterion for defining when one preference structure can be said to be preferred to another, this note has suggested that there is a meaningful sense in which comparability is feasible, at least conceptually.

[7]"Shaping" utility functions is, thus, different from using instruments of the law to regulate conduct. On the issue of the normatively appropriate role of the law in the enforcement of "moral" values (which are often interpretable as involving external effects) see the controversy between Lord Patrick Devlin and Professor H L.A Hart (which builds on the earlier work of John Stuart Mill). (I am indebted to an anonymous referee for calling my attention to this literature.) Papers by Devlin, Hart, and others who participated in the continuing debate appear in a volume edited by Richard A. Wasserstrom

REFERENCES

K. J. Arrow, "Social Responsibility and Economic Efficiency," *Publ. Pol.*, Summer 1973, *21*, 303–17.

R. McKean, "Spillovers from the Rising Value of Time," *Quart. J. Econ.*, Nov. 1973, *87*, 638–40.

M. McManus, "Social Welfare Optimization With Tastes As Variables," disc. paper, Univ. Birmingham, England, Jan. 1976.

John Rawls, *A Theory of Justice*, Cambridge, Mass., 1971.

A. Sen, "Behaviour and the Concept of Preference," *Economica*, August 1973, *40*, 241–59.

Richard A. Wasserstrom, *Morality and the Law*, Belmont 1971.

C. Weizsäcker, "Notes on Endogenous Change of Tastes," *J. Econ. Theory*, Dec. 1971, *3*, 345–72.

The Value of Time in Consumption and Residential Location in an Urban Setting

By ODED HOCHMAN AND HAIM OFEK*

Several economists have recently studied the location decisions of different population groups in towns where these groups are often classified according to income. In the context of a partial equilibrium model, for example, Richard Muth has argued that if housing is a superior good, high wage earners will live farther away from urban centers than will low wage earners; and his empirical evidence supports both the assumption and the conclusion.[1] Margaret Reid and John Kain provide results corroborating, respectively, an income elasticity substantially greater than one and the predicted residential distribution of income groups. Using a specific general equilibrium model, Edwin Mills found that the location pattern was consistent with that of Muth, when the income elasticity exceeded one and was indeterminate when it fell short of this figure. Finally, although Martin Beckmann and A. Montesano assume away the income elasticity problem, their findings are not inconsistent with Muth's theory.

However, several recent empirical studies[2] have estimated the income elasticity of housing to average about 0.7, a finding inconsistent with Muth's theory and one which opens up the whole question of relative residential location once again.

It is accordingly the purpose of this paper to amend the Muthian theory reestablishing congruence between the model's predictions and the empirical findings. The fundamental amendment involves treatment of the time constraint in the framework of consumer choice. The value of time as an important cost of commuting and therefore an ingredient in the mechanism of residential choice is generally both recognized and considered in analyses of location decisions. However, the value of time may also affect residential decisions through its role in the cost of consumption activities other than commuting. By neglecting the tradeoff between housing and time in consumption net of commuting, most studies have overemphasized the tradeoff between housing and commuting costs. By considering both tradeoffs, the present paper works both to correct the previous imbalance in the theory's treatment and to indicate that under fairly plausible conditions high wage earners may still choose to reside farther away from urban centers even if housing is an inferior good.

I. Individual Choice and Constraints

Consider a typical individual assumed to maximize a utility function in two normal goods, H and Z, of the form

$$U = U(H, Z) \qquad (1)$$

where H is residential land as a proxy for housing (see Alonso, Muth (1969), and Beckmann), and Z is a composite commodity. The household produces Z by combining market goods (excluding land) x, and time in consumption t (see Becker), in

*Visiting assistant professor of economics, Princeton University; and lecturer in economics, the Hebrew University, currently visiting Columbia University. This paper was written while we were research fellows, department of economics, University of Chicago, financially supported by National Science Foundation Grant ENV75-23397, and by the Rockefeller and Sloan Foundations. We are indebted to G. Becker, J. Heckman, E. Helpman, P. Linneman, D. Pines, M. Reid, T. W. Schultz, G. Stigler, and a referee for helpful comments and suggestions. None of the aforementioned are responsible for the views expressed in this paper or any remaining mistakes.

[1]Further evidence supporting the direct income-distance relation is provided in the Muth (1969) investigation of six cities. The empirical findings indicate that the simple regression coefficients of income on distance are positive and significantly greater than zero (at the 0.1 level) for all six cities.

[2]See, for example, John Campbell and Barton Smith, Geoffrey Carliner, and R.K. Wilkenson.

accordance with a linear homogeneous production function

$$(2) \qquad Z = F(x, t)$$

The respective cost function is given by

$$(3) \qquad C = C(Z; p, w)$$

where p is the price level of market goods and w is the wage rate (the price of time). The shadow price of the commodity Z is given by

$$(4) \qquad \pi = \frac{\partial C}{\partial Z}(Z; p, w)$$

In all that follows, p is assumed to be fixed. Hence, by the linear homogeneity of (2), π is a monotonically increasing function in w

$$(5) \qquad \pi = \pi(w) \qquad \partial\pi/\partial w \geq 0$$

Moreover, since $\partial C/\partial Z = C/Z$

$$(6) \qquad \pi \cdot Z = C(Z; p, w)$$

With the further standard simplifying assumption (see Muth, 1969, Beckmann, and Mills) that the only cost factor in transportation is commuting time, the budget constraint is given by

$$(7) \quad RH + C(Z; p, w) = w(\mathrm{T} - D) + V$$

where R is rent per unit of land, T is total time available to the household, V is nonearned income, and D is time spent in commuting. Normalizing on T (T = 1), and substituting in (6), (7) can be rewritten

$$(7') \quad RH + \pi Z = w(1 - D) + V$$

The term D stands, therefore, for the proportion of time spent in commuting and may also be interpreted as a measure of the distance between the residential location and the Central Business District (CBD).

II. Equilibrium in the Residential Ring

Consider the environment as a system of cities in equilibrium with free and costless flow of population between cities. Consider also two groups of population, $i = 1, 2$, with the same utility function of type (1) above, but with different wage rates and nonlabor incomes. Thus, the overall equilibrium utility levels in the system will be

$$(8) \qquad U(H_i, Z_i) = u_i \qquad i = 1, 2$$

where u_i, the utility level of group i, is a parameter for all members in each population group at equilibrium.

Thus, the utility function reduces to a single indifference curve for each type of population. Note that equal utility levels between cities does not necessarily imply equal wage rates. Without restrictions on the generality, further assume that

$$(9) \qquad u_1 > u_2$$

Consider now a city in this system containing both types of population, and assumed as is standard, to be circular with a CBD of given radius ϵ where the labor market is located.

Each population group has its own bid-rent function, that is, the maximum rent an individual from that group is ready to pay in each location. Competitive equilibrium implies that the group residing in a given location is the one with the higher bid rent (see Alonso and Mills).

Let $R_i(D)$ be the bid-rent function for population group i. For every D, $R_i(D)$ is determined by the solution to the problem

$$(10) \qquad R_i(D) = \operatorname*{Max}_{Z, H} R_i \qquad i = 1, 2$$

subject to (7′) and (8). The first-order condition necessary for that maximization is

$$(11) \qquad R_i = \frac{U_H}{U_Z}\pi_i \qquad i = 1, 2$$

The properties of the bid-rent functions can be worked out as follows. Differentiating the budget constraint (7′) with respect to distance D results in

$$(12) \quad H_i R_i' + R_i H_i' + \pi_i Z_i' = -w_i \qquad i = 1, 2$$

where a prime above a letter denotes a partial derivative with respect to distance D. After differentiating totally (8) and then substituting in (11), the result is

$$(13) \qquad R_i H_i' + \pi_i Z_i' = 0 \qquad i = 1, 2$$

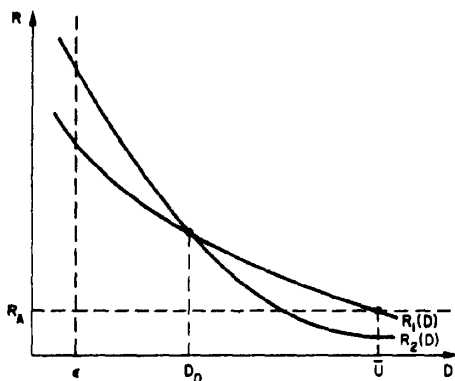



FIGURE 1

Solving from (12) and (13) for R_i',

$$(14) \qquad R_i' = -w_i/H_i < 0 \qquad i = 1,2$$

reflecting the well-known spatial equilibrium condition $R_i'H_i + W_i = 0$.[3] This condition indicates that a household is in equilibrium when moving one unit of distance away from the *CBD* increases commuting cost by exactly the same amount it reduces housing cost. By further differentiating (14), noting that $H' > 0$,

$$(15) \qquad R_i'' = \frac{w_i}{H_i} H_i' > 0 \qquad i = 1,2$$

Equations (14) and (15) indicate that the bid-rent curves must be concave and decreasing. Two typical bid-rent curves obeying these properties are shown in Figure 1.

III. Relative Location Due to Difference in Nonearned Income

Investigation of relative residential location due to differences in nonearned income can now proceed. Consider two population groups with different levels of nonearned income, but equal in all other respects. Thus

$$(16) \quad w_1 = w_2 = w \quad \text{but} \quad V_1 > V_2$$

The presence of two population groups at equilibrium in the same city implies the existence of at least one intersection point between the two respective bid-rent functions. Consider then, such an intersection point D_0 (see Figure 1), where $R_1(D_0) = R_2(D_0)$. By (5) and (16) $\pi_1 = \pi_2$ and since H is a normal good $H_1 > H_2$. Hence, substitution of these relations into (14) obtains

$$(17) \qquad |R_1'| = w/H_1 < w/H_2 = |R_2'|$$

Namely, the bid-rent function of the poorer group is steeper in the neighborhood of the given intersection point. Since this particular intersection point was arbitrarily chosen, this result must obtain in the neighborhood of any intersection point. Two outcomes follow: first, the bid-rent curves of the two population groups have one and only one intersection point; second, $R_1 < R_2$ for $\epsilon \leq D < D_0$, but $R_1 > R_2$ for $D_0 < D \leq \bar{U}$, where D_0 and $\bar{U}$ are the intersection point and the boundary of the town, respectively (see Figure 1).

The first outcome indicates that a well-defined and unique dividing line between an inner residential ring and an outer residential ring has been established in the framework of the given assumptions. The second outcome indicates that the bid rents of the wealthy dominate that market in the outer ring. Hence, this result follows:[4]

COROLLARY 1: *Two local groups of population with different levels of nonearned income, but equal in all other respects, would tend to locate in the following order· the poorer group closer to the CBD, the wealthier farther away.*

IV. Relative Location Due to Wage Differentials

Consider now two local groups of pure wage earners (with no other income aside from earnings) facing different wage rates, but equal in all other respects. In particular,

$$(18) \quad w_1 > w_2 \quad \text{but} \quad V_1 = V_2 = 0$$

Substitution into (7') of $V_i = 0$, followed by division by H_i and substitution for w/π

[3] See, for example, Muth (1969) and Mills.

[4] This result is consistent with earlier findings of Becker and Muth (1969).

from (14), results in

$$R_i + \pi_i \frac{Z_i}{H_i} = -R_i'(1 - D) \quad i = 1,2 \tag{19}$$

Since $R_i' < 0$, $-R_i' = |R_i'|$. Hence, by subtracting (19) for $i = 2$ from (19) for $i = 1$,

$$(|R_1'| - |R_2'|)(1 - D) = (R_1 - R_2) + \left(\frac{\pi_1 Z_1}{H_1} - \frac{\pi_2 Z_2}{H_2}\right) \tag{20}$$

At an intersection point D_0, at which $R_1(D_0) = R_2(D_0) = R$, equation (20) can be expressed in a slightly different form:

$$(|R_1'| - |R_2'|)(1 - D) = R(1/S_{H_1} - 1/S_{H_2}) \tag{21}$$

where S_{H_i} is the share of housing in total expenditure on consumption, that is,

$$S_{H_i} = \frac{R_i H_i}{R_i H_i + \pi_i Z_i} = \frac{R_i H_i}{w_i(1 - D)} \quad i = 1,2 \tag{22}$$

From (21), it is clear that

$$|R_1'| \gtreqless |R_2'| \quad \text{as} \quad S_{H_1} \lesseqgtr S_{H_2} \tag{23}$$

If $S_{H_1} > S_{H_2}$ everywhere, then for $\epsilon \leq D < D_0$ (i.e., everywhere in the inner ring), $R_1 < R_2$, and for $D_0 < D \leq \bar{U}$ (i.e., everywhere in the outer ring), $R_1 > R_2$. If $S_{H_1} < S_{H_2}$, these inequalities are reversed. If $S_{H_1} = S_{H_2}$ everywhere, the two bid-rent curves coincide and $R_1 = R_2$ everywhere. Note also that as long as $S_{H_1} \neq S_{H_2}$ everywhere, D_0 must be the only intersection point.

A preliminary conclusion with respect to the residential distribution of wage earners then follows:

COROLLARY 2: *Two local groups of pure wage earners with different wage rates, but equal in all other respects, would tend to locate in the following order: the group with the lower share of housing at each given level of rent will locate closer to the CBD; the group with the higher share will locate farther away. Equal housing shares, on the other hand, imply a mixed residential distribution of the two groups.*

Note that this is a refutable hypothesis for which direct data might actually exist. Indeed, the whole issue boils down to the question of which group tends to consume a larger share of housing, and the following analysis attempts to characterize the above result in more familiar terms.

The second term on the right-hand side of (21) can be rewritten as follows:

$$1/S_{H_1} - 1/S_{H2} = \left[1/S_{H_1} - (1/S_H)_{\substack{\pi=\pi_1\\ U=u_2}}\right] + \left[(1/S_H)_{\substack{\pi=\pi_1\\ U=u_2}} - 1/S_{H_2}\right] \tag{24}$$

Defining the bracketed terms on the right-hand side of equation (24) as ΔI (the income effect), and Δs (the substitution effect), respectively, allows equation (21) to be rewritten:

$$(|R_1'| - |R_2'|)(1 - D) = R(\Delta I + \Delta s) \tag{25}$$

The sign of ΔI is determined by the standard definition of the income elasticity as follows:

$$\Delta I \lesseqgtr 0 \quad \text{as} \quad \eta \gtreqless 1 \tag{26}$$

where η is the income elasticity of H. Similarly, when $\pi_1 > \pi_2$, the sign of Δs is determined as follows:

$$\Delta s \lesseqgtr 0 \quad \text{as} \quad \left.\frac{\partial S_H}{\partial \pi}\right|_{\substack{dU=0\\ dR=0}} \gtreqless 0 \tag{27}$$

When $\pi_1 = \pi_2$, $\Delta s = 0$.

Let σ stand for the Allen-Hicks elasticity of substitution of U, that is,

$$\sigma = \left.\frac{d \log Z/H}{d \log R/\pi}\right|_{dU=0} = 1 + \left.\frac{d \log S_Z/S_H}{d \log R/\pi}\right|_{dU=0} \tag{28}$$

Since $S_H + S_Z = 1$, solving for S_H in terms of S_Z/S_H obtains

$$S_H = \frac{1}{S_Z/S_H + 1} \tag{29}$$

Differentiating both sides of (29) with respect to π results in[5]

$$\left.\frac{\partial S_H}{\partial \pi}\right|_{\substack{dU=0\\ dR=0}} = \frac{S_H S_Z}{\pi}(\sigma - 1) \tag{30}$$

Since all the variables on the right-hand side of equation (30) are positive, the sign of the expression depends on whether $\sigma \gtreqless 1$. Hence by (27) and (30)

$$\Delta s \begin{cases} = 0 & \text{iff} \quad \sigma = 1 \quad \text{or} \quad \pi_1 = \pi_2 \\ > 0 & \text{iff} \quad \sigma < 1 \\ < 0 & \text{iff} \quad \sigma > 1 \end{cases} \tag{31}$$

Table 1 sums up the results derived from equation (26). Corollary 3 follows immediately.

COROLLARY 3: *Two groups of pure wage earners residing in the same city, differing in their wage rates but equal in all other respects, would tend to locate in the following way:*

a. *High (low) wage earners will reside farther away from (closer to) the CBD than the low (high) wage earners if both the income elasticity of housing is greater than one ($\eta > 1$) and the elasticity of substitution between housing and all other goods is greater than one ($\sigma > 1$).*

b. *High (low) wage earners will reside closer to (farther away from) the center of town than low (high) wage earners if both $\eta < 1$ and $\sigma < 1$.*

c. *If $\eta = 1$ and either $\sigma = 1$ or $\pi_1 = \pi_2$, a mixed residential distribution of the two groups will result.*

TABLE 1

	Sign of $(\lvert R_1' \rvert - \lvert R_2' \rvert)$		
	$\eta < 1$	$\eta = 1$	$\eta > 1$
$\sigma < 1$	+	+	?
$\sigma = 1$	+	0	–
$\sigma > 1$	?	–	–

d. *If either $\eta > 1 > \sigma$ or $\eta < 1 < \sigma$, the the final outcome is indeterminant on th basis of qualitative information alone. How ever for every $\sigma_0 < 1$, there exists $\eta(\sigma_0) > 1$; $\partial\eta(\sigma_0)/\partial\sigma_0 < 0$ so that if η $\eta(\sigma_0)$, the high wage earners will live farthe away from the center than low wage earner. In the same way for every $\eta_0 < 1$, ther exists a $\sigma(\eta_0) > 1$; $\partial\sigma(\eta_0)/\partial\nu_0 < 0$, so th $\sigma > \sigma(\eta_0)$, then again high wage earners wi reside farther away than low wage earners.*

V. Conditions under which the Substitution and Income Effects Hold

Recent literature dealing with the relati location of different income groups h taken into account only the income effec Muth and Mills, for instance, by not in cluding time in consumption have implicitl assumed $\pi_1 = \pi_2$. Their results, reflectin situations for which the income elasticity is the only determining factor, are con sistent therefore with the three entries alon the second row in Table 1. Beckmann an Montesano have assumed in addition utility function with $\eta = 1$, and thus the results are equivalent to the special indi ference case as stated in part c of Corol lary 3 above (or the zero entry in Table 1).

We may consider the wage effect to be measure of the market productivity of th worker. If when market productivity in creases, home productivity increases as wel then π, the cost of the consumption com modity, may not change with w. In thi case, since $\partial\pi/\partial w = 0$, the substitutio effect vanishes and the income effect is th sole determinant of residential locatior This result is consistent with the tradition approach as argued by Muth and by Mill In this case, high wage earners will locate i

[5]The following steps are involved in the derivation of the right-hand side of (30)

$$\frac{d}{d\pi}\frac{1}{S_Z/S_H + 1} = -\frac{d(S_Z/S_H)}{d\pi}(S_Z/S_H + 1)^{-2} = S_H^2\frac{d(S_Z/S_H)}{d(R/\pi)}\frac{d(R/\pi)}{d\pi} = \frac{S_H S_Z}{\pi}\frac{d\log(S_Z/S_H)}{d\log(R/\pi)}$$

By substituting in the above equation $(\sigma - 1)$ from equation (28), we get (30).

the outer ring if the income elasticity of housing η is greater than unity, and will locate in the inner residential ring if η is less than unity.

The relationship between market and home productivity depends upon the causal structure of those differences. According to Robert Michael, a high correlation between the two productivities is expected if the difference in wage rates reflects personal ability or general training and schooling. Market-specific training, discrimination, or random variation of wages, however, would lower the correlation and bring about strong substitution effects.

VI. Adjustments in Response to Further Household Characteristics

The formulation of the model is now extended to cover differences in household size and composition. The budget constraint (7′) is generalized as follows:

$$(32)\quad RH + \pi Z = n(1 - D)w + mw' + cw'' + v$$

where n is the number of family members in the labor force; w their wage rate; m is adult members not in the labor force; w' is the unit value of their time valued as a shadow price; and c, the number of children present in the household with w'' the unit value of their time. The variables n, m, c, w, w', w'', may be viewed as vectors if further breakdowns are desired.

To see the modification implied by this generalization in our main results, substitute (32) for (7′) and proceed with the same analysis, and calculations as in the simple model (Section III). Instead of (14), the generalized version of this equilibrium condition becomes

$$(33)\quad R' = -nw/H$$

By substituting (33) into (32) and rearranging,

$$(34)\quad -R' = R/(\alpha - D)S_H$$

where

$$(35)\quad \alpha = 1 + mw'/nw + cw''/nw + V/nw$$

Comparing the bid-rent functions of two population groups ($i = 1,2$) at the intersection point ($R_1 = R_2 = R$ and $D_1 = D_2 = D$),

$$(36)\quad |R_1'| - |R_2'| = R\left[\frac{1}{(\alpha_1 - D)S_{H_1}} - \frac{1}{(\alpha_2 - D)S_{H_2}}\right]$$

Equation (36) is a generalization of (21) allowing for differences in family composition and labor force participation of its members. Without such differences, $\alpha_1 = \alpha_2 = 1$ and all the results obtained in Corollaries 2 and 3 follow straightforwardly. If differences in family characteristics do exist, $\alpha_1 \neq \alpha_2$, and further implications arise.

For example, let us consider the effect of a larger number of earners in the household on its relative location. Consider a typical case of two groups ($i = 1,2$) of husband-wife families assumed to be identical in all respects except that in group one ($i = 1$), two members are employed in the *CBD*, whereas in the second group ($i = 2$) only one is in the labor force.

From (35)

$$(37)\quad \alpha_1 = 1 + \frac{c_0 w''}{2w} < 1 + \frac{w'}{w} + \frac{w''}{w} c_0 = \alpha_2$$

Substitution of (37) into (36) results in the following corollary:[6]

[6]Let D be the intersection point of the two bid rents. If both households have exactly the same characteristics, they must be on the same indifference curve. This implies that $w(1 - D) < w' < w$ so that $\pi_1 > \pi_2$ Equation (36) at D can be written as follows

$$|R_1'| - |R_2'| = R\left[\frac{1}{(\alpha_1 - D)}\left(\frac{1}{S_{H_1}} - \frac{1}{S_{H_2}}\right) + \frac{1}{S_{H_2}}\left(\frac{\alpha_2 - \alpha_1}{(\alpha_1 - D)(\alpha_2 - D)}\right)\right]$$

$$= R\left(\frac{1}{\alpha_1 - D}\Delta s + \delta\right)$$

where $\delta = \frac{1}{S_{H_2}}\left(\frac{\alpha_2 - \alpha_1}{(\alpha_1 - D)(\alpha_2 - D)}\right) > 0$

Δs depends on the elasticity of substitution σ alone: the income elasticity of housing η has no effect at all. Corollary 4′ can now replace Corollary 4:

COROLLARY 4′: *Let us consider two households iden-*

COROLLARY 4: *Households of working wives will reside closer to the CBD than households of nonworking wives, unless the share of housing consumed by the former is markedly larger than the share consumed by the latter.*

Corollary 5 follows in much the same way.[7]

COROLLARY 5: *Larger families will reside farther away from the center than smaller ones, unless the share of housing consumed by the latter is markedly larger than the share consumed by the former.*

In practice, the male is the provider in most single earner families. Therefore, the implications of Corollary 4 are that if shares are similar, the proportion of women employed will be higher among those residing near the *CBD* than those farther away. Indeed, empirical findings by Kain indicate that higher proportions of female *CBD* workers in Detroit (1953) resided in nearby residential rings than did the proportions of male *CBD* workers. Similarly, the findings by Rees and Shultz indicate that the three predominantly female occupations show the shortest mean distance traveled to work among twelve selected occupations in Chicago (1963).

The empirical findings of both Kain and Muth directly corroborate Corollary 5. A positive correlation exists between family size and the distance of residences from the *CBD*.

The intuitive appeal of the findings in Corollaries 4 and 5 is also apparent. There is a tradeoff between commuting costs and benefits from housing services, and while the costs of added distance from the *CBD* are related basically only to number of wage earners, the benefits from additional housing services if shares are similar, are related to total family size. The existence of sizeable moving costs, however, indicates that households will make their location decisions according to planned family size in some foreseeable future rather than the *actual* number at any given moment.

Equation (35) indicates that a higher imputed price of children time w'' will result in a greater α, so that distance from center and w'' are positively related. Because the value of childrens' time tends to increase with age,[8] the theory also suggests that households with older siblings would live farther away— unless sizeable moving costs again retard or preclude such movement.

Finally, because differences in intergenerational preferences will produce permanent differences in the imputed childrens' value of time w'', families with a higher preference for child quality would tend to reside farther away from the center. In fact, suburbs have better educational services than their inner cities. Traditionally family migration to the suburbs has been explained by the desire for good education for children. The causation suggested by our argument clearly runs in the opposite direction it ascribes the better educational services to the demand for such services by the typical population that would reside in the suburbs anyway.

tical in all characteristics, except that one has two earners and the other only one. Then there exists a $\sigma_0 > 1$ so that if σ, the elasticity of substitution, is smaller than σ_0, the family with two earners will reside closer to the CBD If $\sigma > \sigma_0$ this family will reside farther away from the center If $\sigma = \sigma_0$, both households will be indifferent about their relative locations

[7]In the cases where there are differences in household size, the conditions for relative locations cannot be described in terms of elasticities, because the household consumption production functions then differ. The difference in the respective prices of Z cannot be determined.

[8]The major use of children's time in modern societies, particularly in urban areas where child employment is very limited, is in producing human capital The productivity of a person in this activity is believed to be positively correlated at each given point of time with previously accumulated human capital (see Yoram Ben-Porath). In this respect, families would tend to value the time of older children more than that of younger offspring.

REFERENCES

William Alonso, *Location and Land Use*, Cambridge, Mass. 1964.

G. S. Becker, "A Theory of the Allocation of

Time," *Econ. J.*, Sept. 1965, *75*, 493-517.

M. J. Beckmann, "On the Distribution of Urban Rent and Residential Density," *J. Econ. Theory*, June 1969, *1*, 60-67.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earning," *J. Polit. Econ.*, Aug. 1967, *75*, 352-65.

J. Campbell and B. Smith, "Aggregation Bias and the Demand for Housing," mimeo, 1975.

G. Carliner, "Income Elasticity of Housing Demand," *Rev. Econ. Statist.*, Nov. 1973, *15*, 528-32.

J. F. Kain, "The Journey to Work as a Determinant of Residential Location," *Papers Proc. Reg. Sci. Assn.*, *9*, 1962, 137 60.

R. T. Michael, "Education in Nonmarket Production," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, 306-27.

Edwin S. Mills, *Urban Economics*, Glenview 1972.

A. Montesano, "A Restatement of Beckmann's Model on the Distribution of Urban Rent Residential Density," *J. Econ. Theory*, Apr. 1972, *4*, 329-54.

Richard F. Muth, *Cities and Housing*, Chicago 1969.

———, *Urban Economic Problems*, New York 1972.

Albert E. Rees and George P. Schultz, *Workers and Wages in an Urban Labor Market*, Chicago 1970.

Margaret G. Reid, *Housing and Income*, Chicago 1962.

R. K. Wilkenson, "The Income Elasticity of Demand for Housing," *Oxford Econ. Pap.*, Nov. 1973, *25*, 361-77.

Labor Supply and the Payroll Tax: Note

By ROBERT A. MOFFITT*

In a recent paper in this *Review*, Duncan MacRae and Elizabeth MacRae perceived an important aspect of the labor-supply effects of the payroll tax—namely, that the effect is crucially dependent upon the individual response to a "kinked" budget constraint. However, their actual conclusions on the direction of the response need major qualification. In particular, they assumed that all responses would be marginal, when in fact the possibility of nonmarginal response is always present when non-linear budget constraints shift. This latter possibility makes the labor-supply response much more ambiguous.

This is illustrated in Figure 1 (modeled after Figure 1 of MacRae and MacRae). The budget constraint before the imposition of the tax is *AB*, giving the utility maximizer a choice of any income-leisure combination along *AB*. After the imposition of the tax—composed of a marginal tax rate on earnings up to a maximum level of earnings E—the (disposable income) budget constraint is *BDC*. (Y_n is the amount of nonlabor income.) At point *D*, the taxable maximum is reached. MacRae and MacRae assumed that an individual initially located above the maximum level (say, point I_1) would relocate along *CD*, resulting in an increase in labor supply if leisure is a normal good (the income effect); and that an individual below the maximum level (say, point I_2) would relocate along *BD*, resulting in a decrease in labor supply if the substitution effect dominates the income effect. However, although this would occur for marginal changes, nonmarginal changes are also possible: an individual initially at I_1 could relocate along *BD*, reducing labor supply; and an individual initially at I_2 could relocate along *CD*, increasing labor supply. Both are possible and not in conflict with any of the assumptions (that leisure is a normal good or that the substitution effect dominates the income effect). Heuristically, imagine a shift from *AB* to *CH*, causing the individual at I_1 to definitely move to *CD*; a subsequent "pivot" of *DH* to *DB* could easily induce the person to relocate along *DB*. Or, imagine a shift from *AB* to *BG*, causing the individual at I_2 to definitely move to *DB* (under the assumptions); a subsequent pivot of *DG* to *DC* could easily induce the person to relocate along *DC*. Thus, the possibility of nonmarginal movements introduces more ambiguity than realized before.[1]

As a sidelight, note that the labor-supply effects of an increase in the taxable maximum and of an increase in the tax rate—policy alternatives that are currently being considered to raise revenue—are both ambiguous for the same reasons. Figure 2 shows an increase in the tax rate as causing a shift from *BDC* to *BD'C'*. Those initially below the maximum level may decrease labor supply (marginally) or increase it (nonmarginally), while those above the maximum level may increase labor supply (marginally) or decrease it (nonmarginally) Figure 3 shows an increase in the taxable earnings maximum from *E* to *E'* as causing a shift from *BDC* to *BD'C'*. Although there is no response by those below the maximum (by revealed preference), those above may either increase labor supply (marginally) or decrease it (nonmarginally). The latter effect is likely to be stronger for this group than

*Economist, Mathematica Policy Research. The comments of an anonymous referee are appreciated. This research was supported by the Department of Health, Education, and Welfare under Contract HEW-100-76-0073. The opinions expressed are those of the author and do not necessarily represent the views of the sponsor.

[1] Empirically, nonmarginal responses are especially likely for working wives and other secondary workers, whose hours worked tend to be quite sensitive to these types of changes Nonmarginal responses are probably less likely for prime-age males, but so are marginal responses.

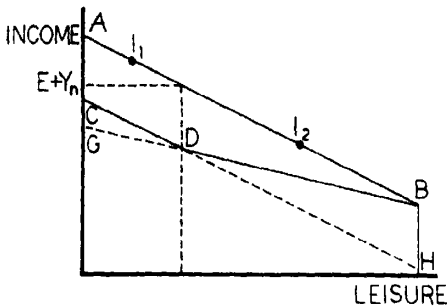



FIGURE 1

FIGURE 3

in Figure 2, since the below-maximum segment has not fallen, but the direction of the net difference between the two cases depends upon the net direction of the response of those below the maximum in the tax-increase case.

The uncertainty of labor-supply choice in the face of non-linear budget constraints occurs not only in the case of the payroll tax, but also in the cases of the federal income tax, many state income taxes, and virtually all income-maintenance programs. In the income-maintenance literature, the analytic problem has already been addressed by Robert Hall, Jonathan Kesselman, and Samuel A. Rea, Jr., who have shown that the labor-supply response to a change in a non-linear budget constraint can be determined if either the utility function or true income and substitution effects are known. However, the formidable empirical problem of estimating either utility function parameters or income and substitution effects with data on workers actually facing non-linear budget lines has yet to be solved (see works by the author and Terence Wales for such attempts). However, in the literature on the incentive effects of the payroll tax, it does not appear that the analytic problem of non-linear budget lines has been recognized to the same extent, as indicated by the paper by MacRae and MacRae, and by other works on the same subject (see, for example, the article by Edgar Browning).

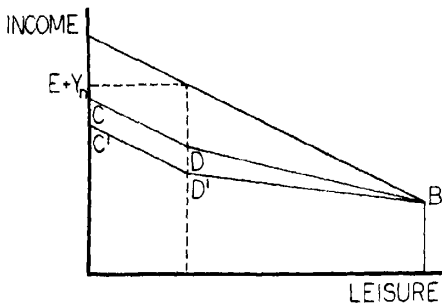
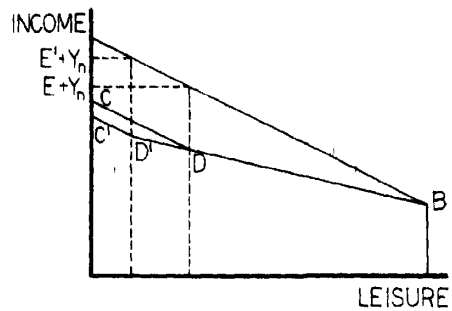



FIGURE 2

REFERENCES

E. K. Browning, "Labor Supply Distortions of Social Security," *Southern Econ. J.*, Oct. 1975, *42*, 243–52.

R. E. Hall, "Effects of the Experimental Negative Income Tax on Labor Supply," in Joseph A. Pechman and P. M. Timpane, eds., *Work Incentives and Income Guarantees*, Washington 1975.

J. Kesselman, "Conditional Subsidies in Income Maintenance," *Western Econ. J.*, Mar. 1971, *9*, 1–20.

C. D. MacRae and E. C. MacRae, "Labor Supply and the Payroll Tax," *Amer. Econ. Rev.*, June 1976, *66*, 408–09.

R. A. Moffitt, "Labor Supply, Kinked Budget Constraints, and the Negative Income Tax," work. paper, Mathematica Pol. Res. 1977.

S. A. Rea, Jr., "Incentive Effects of Alternative Negative Income Tax Plans," *J. Publ. Econ.*, Aug. 1974, *3*, 237–49.

T. Wales, "Estimation of a Labor Supply Curve for Self-Employed Business Proprietors," *Int. Econ. Rev.*, Feb. 1973, *14*, 69–80.

Peak Load Pricing with Stochastic Demand

By Dennis W. Carlton*

Recent articles (see Gardner Brown, Jr. and M. Bruce Johnson, and Michael Visscher) in this *Review* have analyzed peak load pricing under uncertainty. The problem is to choose an output capacity and a price before the realization of stochastic demand. This scenario realistically describes how many firms in both the private and public sector operate. Familiar examples include retail stores, manufacturing firms, restaurants, hotels, airlines, parks, and public utilities. Consistent findings of these models are that the socially optimal price (defined as the price that maximizes expected surplus) should not exceed long-run marginal costs, and that expected profits should be negative. In the optimum, firms need to receive a subsidy to remain in operation. In this paper, I show that when uncertainty enters the demand curve in a multiplicative fashion, then under plausible conditions, these findings are completely reversed. In the optimum, price should exceed long-run marginal costs, and expected profits will be nonnegative. In this optimum, no subsidization of firms is required. Finally, I comment on the appropriateness of using expected surplus as a criterion of social welfare and on the desirability of charging different prices to consumers.

In peak load models, it is necessary to specify the stochastic structure of demand and the rationing scheme that results when there is excess demand. Brown and Johnson consider multiplicative and additive demand uncertainty. Their rationing scheme is one in which individuals with the highest willingness to pay wind up obtaining the goods. Visscher criticized this rationing scheme as being unrealistic. If prices do not clear markets, and recontracting markets do not exist, it does not follow that a rationing scheme will be able to provide goods to those with the greatest willingness to pay. Moreover, if individuals have per capita demand curves, this rationing scheme would imply that an individual would only be able to satisfy his high consumer surplus demand. Visscher examined two different rationing schemes, one where demand is satisfied randomly and the other where those with the lowest willingness to pay wind up being served first. The first scheme might apply to a person trying to get a seat on an airplane. The second could apply to individuals waiting in line for tickets. If those with low value of time get to the line first, and have a low willingness to pay, then we obtain the second rationing scheme. In Visscher's models, uncertainty enters the demand curve additively.

The cost structure in these models is one of constant operating cost b, and constant capacity cost β. Both Brown-Johnson and Visscher reach the conclusion that to maximize expected surplus, the optimal policy is to charge a price p that is less than or equal to the long-run marginal cost $b + \beta$. This means that the firm will necessarily make negative expected profits and therefore require a subsidy to operate. (Since there is a positive probability that some capacity will not be used, a price equal to $b + \beta$ will lead to negative expected profits.)

This paper considers the two rationing schemes of random rationing, and rationing to those with the lowest willingness to pay for the case of multiplicative demand uncertainty.[1] The social welfare function is initially taken to be expected surplus. For the random rationing scheme, the optimal price *exceeds* $b + \beta$, and expected profits equal zero. The firm requires no subsidy to remain in business. In fact, provided there are

*Assistant professor, department of economics, University of Chicago. I thank John Panzar for helpful comments.

[1] As mentioned above, the third rationing scheme (rationing to those with the highest willingness to pay) was examined by Brown and Johnson for the case of multiplicative demand uncertainty.

firms who compete with each other, it is possible that the market can be relied on to achieve this equilibrium. In the rationing to lowest value users, the optimal price again exceeds $b + \beta$, but profits are now positive. A competitive market can achieve this desired price and output policy if taxation is used. Finally, it is argued that expected surplus may be a very poor indicator of social welfare, and that charging a single price to all consumers may be nonoptimal. When uncertainty about obtaining a product arises, consumers' preferences for risk matter. Consumers will trade off price and probability of obtaining the good. Only by pure chance will expected surplus capture these tradeoffs. For the case of identical consumers and the random rationing scheme, I state the conditions under which subsidization or taxation will occur in the socially optimal solution. With customers of different demand riskiness, it is not in general optimal to charge the same price to all.

I. Random Rationing

In this section, it is shown that with random rationing, the price that maximizes expected surplus to society exceeds $b + \beta$, and that at this price expected profits are zero. Let demand be $D(p) = x(p)u$, where p is price, $x(p)$ is the nonstochastic component of demand, and u is a positive random variable with cumulative density $F(u)$. Assume that $x'(p) < 0$. One possible way to think of demand is that u measures the random number of customers while $x(p)$ is the per capita demand curve. Let $x^{-1}(q)$ be the inverse function of $x(p)$. Let $z \equiv sx(p)$ be the capacity chosen. If $D(p) > z$, or equivalently $u > s$, then rationing occurs. Under the random rationing scheme, each unit of demand has an equal chance of being satisfied. The random rationing scheme makes most sense in the case where all consumers have the same quantity demand for the product (for example, consumers demand just one seat on an airplane). (Consumers could, of course, differ in the price they are willing to pay for the product.) In this situation, random rationing implies that all consumers have an equal chance of being satisfied in their demands. Under the random rationing scheme, the total expected surplus to society is

$$(1)\quad S = \int_0^s u\left[\int_0^{x(p)} x^{-1}(q)dq - bx(p)\right]dF(u) + \int_s^\infty \frac{s}{u}\, u\left[\int_0^{x(p)} x^{-1}(q)dq - bx(p)\right]dF(u) - \beta sx(p)$$

The first term is expected surplus when no rationing occurs, the second is expected surplus when demand exceeds available capacity, and the last term is the capacity cost. We wish to maximize S with respect to p and s. For purposes of this paper we need consider only $\partial S/\partial p$.[2] Setting $\partial S/\partial p$ to zero, we obtain

$$\int_0^s u[x'(p)(p-b)]dF(u) + \int_s^\infty s[x'(p)(p-b)]dF(u) - \beta sx'(p) = 0$$

or since $x'(p) \neq 0$

$$(2)\quad (p-b)\left[\int_0^s udF(u) + \int_s^\infty sdF(u)\right] - \beta s = 0$$

Introduce the following notation:

$$(3)\quad I \equiv \frac{1}{s}\left[\int_0^s udF(u) + \int_s^\infty sdF(u)\right] = \frac{1}{s}\left[\int_0^s (u-s)dF(u) + s\right] < 1$$

Solving (2) for p and using (3) and the fact that $I > 0$, we obtain the result that the

[2] We make the usual assumption that the second-order conditions for an interior maximum are satisfied. If we wished to determine the optimal capacity and price, we would need to consider the first-order condition $\partial S/\partial s = 0$. It is possible for the optimal capacity to be either above or below that of the riskless world for the situations considered in this paper.

optimal price satisfies

$$p = \beta/l + b > \beta + b$$

Notice also that expected profits are simply

$$(4)\quad \pi = (p - b)x(p) \cdot \left[\int_0^s u dF(u) + s \int_s^\infty dF(u)\right] - \beta s x(p)$$

The first integral in (4) is related to revenue when demand is less than capacity, while the second integral is related to revenue when demand exceeds capacity. Comparing (2) to (4), it is clear that (2) implies that expected profits are zero in the social optimum.

Price exceeds deterministic long-run marginal cost in this optimum. Despite the divergence between price and marginal cost, it pays to have some unused capacity, and to charge a high price. Unused capacity provides the service of insuring that customers have some probability of satisfying their demand. Expected profits are zero even though $p > b + \beta$, since the costs of unused capacity must be recouped.

As mentioned in the introduction, there are competitive type markets with this peak load structure. In such markets, goods have two relevant characteristics, their price and the probability they are available. Firms will compete with each other by offering the best package (price, probability) until their expected profits are driven to zero. Consumers have preferences between these two attributes. If consumer preferences are identical and are adequately represented by expected consumer surplus, then competing firms will reach the optimum described above. For a proof of this result and more details on competition under inflexible prices and stochastic demand, see my referenced papers.

II. Rationing to Those with Low Willingness to Pay

If value of time is correlated with willingness to pay, then a rationing scheme in which those with low willingness to pay wind up being served first may be realistic, especially when queuing is necessary. I show that for this rationing scheme the optimal price exceeds $b + \beta$, and profits are positive.

For this rationing scheme, expected surplus is

$$S = \int_0^s \left[u\int_0^{x(p)} x^{-1}(q)\,dq - bx(p)\right] dF(u) + \int_s^\infty \left[u \int_{x(p)\frac{(u-s)}{u}}^{x(p)} x^{-1}(q)\,dq - b\frac{s}{u}x(p)\right] dF(u) - \beta s x(p)$$

Setting $\partial S/\partial p$ to zero, we obtain

$$\int_0^s u(p - b)x'(p)\,dF(u) + \int_s^\infty upx'(p)\,dF(u) - \int_s^\infty bsx'(p)\,dF(u) - \int_s^\infty u\,\frac{u-s}{u}\,p^* x'(p)\,dF(u) - \beta s x'(p) = 0$$

where p^* is a function of u, s, and p, defined by $x(p^*) = [(u - s)/u]x(p)$. Notice that if $u > s$ then $p^* > p$. Rewriting the above and using the fact that $x'(p) \neq 0$ we obtain

$$\int_0^s u(p - b)\,dF(u) + (p)\int_s^\infty (u - s)\,dF(u) + (p - b)s\int_s^\infty dF(u) - \int_s^\infty (u - s)p^*\,dF(u) - \beta s = 0$$

or

$$(5)\quad \int_0^s u(p - b)\,dF(u) + (p - b)s\int_s^\infty dF(u) - \beta s + \int_s^\infty (u - s)(p - p^*)\,dF(u) = 0$$

The last term in (5) is negative since $p^* > p$ over the range of integration. From (4), it is clear that the first three terms in (5) are profits divided by $x(p)$. Therefore, it follows from (5) that profits are positive in the optimal solution. For positive profits, we must have $p > b + \beta$.

In this optimal solution, firms should be run at a profit, no subsidization is necessary. Since competitive markets compete profits away, achieving the optimal equilibrium competitively using only one price is impossible without some intervention. Taxing away these profits would be one way to achieve this optimum in a competitive market.

III. Consumer Preferences and Different Prices for Consumers

When goods are not always available to consumers, the probability of obtaining the good becomes a characteristic of the good. Consumers will have indifference curves between price and probability of obtaining the good. Calculation of the social optimum should take these preferences into account.[3] Using expected surplus as a measure of social welfare is valid only if the indifference curves of consumers for price and probability coincide with equal expected surplus lines. In general, there is no reason to expect this to occur.

If expected surplus does not measure consumer preferences, what does the social optimum look like? This question is examined in my forthcoming paper for the case of identical consumers, a random rationing scheme, and multiplicative demand uncertainty. The social optimum (i.e., the one which maximizes the expected utility of a typical consumer) will in general involve either subsidization or taxation of the firms. If the marginal utility of income is higher when a greater variety of goods are available, then subsidization occurs in the optimum; under the opposite assumption, taxation of firms occurs. With no restrictions on preferences, it is possible for optimal price to be either above or below deterministic long-run marginal cost.

With different types of consumers, the optimal price depends on interpersonal welfare comparisons and on each individual's riskiness of demand. Since demand uncertainty imposes costs on firms in the form of unused productive capacity, charging different customers different prices according to their riskiness of demand is desirable, whether or not expected surplus is the appropriate welfare criterion. Presumably, the difficulty of identifying different risk classes explains why single price models have received most of the attention in the literature on peak load pricing under uncertainty (Meyer presents a notable exception).

IV. Conclusions

This paper makes two points. With realistic rationing schemes and multiplicative demand, the operating policies which maximize surplus to society involve nonnegative profits and a price *above* long-run marginal costs. This result contrasts sharply with previous findings regarding peak load problems. For many situations, multiplicative demand uncertainty seems more relevant than additive uncertainty. This is one reason why econometric equations are often run in double-*log* form. For the cases examined here, no subsidization of firms need occur in the optimum. If a competitive market is possible, then competition, perhaps combined with taxation, can achieve the social optimum.

The second point emphasizes that using expected surplus as the welfare function and charging the same price to all consumers may both be undesirable in analyzing peak load problems under uncertainty. The probability of obtaining a good is a characteristic of the good for which consumers have preferences. The social optimum should take these preferences into account.

[3]See Robert Meyer for a model which imposes exogenous requirements on the acceptable levels for probabilities, and then optimizes subject to these constraints.

There is no compelling reason why expected surplus should adequately represent these preferences. When customers differ in their riskiness of demand, they impose different costs on a firm and, if possible, it would be desirable to charge customers according to their riskiness of demand.

REFERENCES

G. Brown, Jr. and M. B. Johnson, "Public Utility Pricing and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119–28.

D. Carlton, "Market Behavior with Demand Uncertainty and Price Inflexibility," *Amer. Econ. Rev.*, forthcoming.

———, "Uncertainty, Pricing and Production Lags," *Amer. Econ. Rev. Proc.*, Feb. 1977, *67*, 244–49.

R. Meyer, "Monopoly Pricing and Capacity Choice Under Uncertainty," *Amer. Econ. Rev.*, June 1975, *65*, 326–37.

M. Visscher, "Welfare Maximizing Price and Output with Stochastic Demand: Comment," *Amer. Econ. Rev.*, Mar. 1973, *63*, 224–29.

Search in the Labor Market and the Duration of Unemployment: Note

By ROBERT M. FEINBERG*

In a recent article in this *Review*, John Barron presents an interesting theoretical analysis, extending the model of optimal job search to take account of the time required "... to search firms for vacancies as well as the time it takes to search vacancies for suitable wages" (p. 934). He gives some evidence to show that the number of vacancies in *U.S.* manufacturing v is inversely related to both the average duration of unemployment D and the average probability that a wage offer is accepted P. Moreover, he claims these data to be consistent with the theoretical model, without resorting to an assumption of faulty perceptions.[1] However, Barron's theory can be extended and more strongly confirmed by his data than he indicates; Barron does not obtain a theoretical prediction for dD/dv, and hence is unable to establish a solid link between his theory and his evidence.[2] This note makes clear the assumption required for dD/dv to be negative and investigates the likelihood of this condition being met.

In Barron's model, the job seeker considers both the distribution of wage offers and the probability of receiving a nonzero wage offer (locating a vacancy) in determining his search strategy. A minimum acceptable wage offer (or reservation wage) a is determined by equating the marginal cost of searching one more period with the expected marginal return. This return is equal to the probability of receiving a wage offer in the next period θ, times the expected wage gain if an offer is received. The probability of receiving an *acceptable* wage offer in any period of search is equal to θP, where P is the probability of accepting a wage offer *given that one is received*; hence, average duration of unemployment D is $1/\theta P$.

The theoretical connection between D and the number of vacancies v is obtained by letting θ (the probability of finding a vacancy) equal vk/nq, where k is the number of firms which can be contacted in each period of search, n is the total number of firms, and q is the number of occupations or types of labor in the economy. As v increases, the initial effect is to *increase* the probability of finding a vacancy; however, since the expected marginal return from search is now higher, the job seeker will adjust upwards his reservation wage—reducing his probability of accepting a wage offer. The net effect on the probability of receiving an acceptable wage offer (and, hence, on D) is not immediately obvious.

In his Appendix, Barron derives $\partial D/\partial v$ (this should be dD/dv) $= -(nq/v^2k)(1/P) - (nq/vk)((\partial P)/(\partial v)/P^2)$. Extending his analysis,

$$(1)\quad dD/dv = -(1/P)(nq/vk)\left(1/v + \frac{\partial P}{\partial v}/P\right)$$

In order for dD/dv to be less than zero (as Barron finds his evidence suggesting), it must be the case that

$$(2)\quad 1/v + \frac{\partial P}{\partial v}/P > 0 \rightarrow \frac{\partial P}{\partial v}/P > -1/v \rightarrow \frac{v}{P}\frac{\partial P}{\partial v} > -1$$

That is, the elasticity of P with respect to v must be less than one in absolute value (if negative) for dD/dv to be negative.

*Assistant professor of economics, Pennsylvania State University. I would like to thank William R. Johnson and two anonymous referees for their helpful comments, and to acknowledge financial support from a doctoral dissertation grant (no. 91-51-75-43) from the U.S. Department of Labor, Manpower Administration.

[1]The following analysis, as does Barron's, ignores the possibility of an increase in average duration due to an unperceived decline in wage possibilities. Barron mentions this possibility as "Effect 3" but interprets the data as denying its importance.

[2]That is, an empirical finding suggesting $dD/dv > 0$ could also be consistent with Barron's model *as he presents it*.

For the case of a finite time horizon and no discounting,[3]

$$(3)\quad \frac{\partial P}{\partial v} = \frac{-f(a)}{vP} \int_a^{\infty} (w - a) f(w) dw$$

(obtained by combining Barron's equations (A4), (2), and (3)), where w is the wage offered, distributed with a density function $f(w)$. Then

$$(4)\quad \eta = \frac{v}{P}\frac{\partial P}{\partial v} = -(f(a)/P^2) \int_a^{\infty} (w - a) f(w) dw$$

The remainder of this note will consider two alternative functional forms for $f(w)$, the rectangular (or uniform) and the normal, and show that the necessary condition $\eta > -1$ is always satisfied.[4]

The general rectangular distribution has the following density function:

$$(5)\quad f(w) = \begin{cases} 1/2m; \text{ for } b - m < w < b + m \\ 0; \text{ elsewhere} \end{cases}$$

Since

$$(6)\quad P = \int_a^{\infty} f(w) dw = \int_a^{b+m} \frac{1}{2m} dw = \frac{1}{2m}[b + m - a]$$

we can write $a = b + m - 2mP$. Then

$$(7)\quad \eta = -\left(\frac{f(a)}{P^2}\right) \int_a^{b+m} (w - a) f(w) dw = -\left(\frac{1}{2mP^2}\right)\left[\int_a^{b+m} \frac{1}{2m} w dw - aP\right] = -\frac{1}{2mP^2}\left[\frac{1}{2m}\left(\frac{(b + m)^2 - a^2}{2}\right) - aP\right] = \frac{-(b + m)^2 + a^2 + 4amP}{8m^2P^2} = -\frac{1}{2}$$

If w is normally distributed, $N(\mu, \sigma^2)$, equation (4) becomes

$$(8)\quad \eta = -\frac{1}{P\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2} \cdot \left[\frac{1}{P\sigma\sqrt{2\pi}} \int_a^{\infty} w e^{-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2} dw - a\right]$$

Let $z = (w - \mu)/\sigma$; $w = z\sigma + \mu$; $dz = dw/\sigma$; $dw = \sigma dz$. Then,

$$(9)\quad \int_a^{\infty} w e^{-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2} dw = \int_{\frac{a-\mu}{\sigma}}^{\infty} (z\sigma + \mu) e^{-z^2/2} \sigma dz = \sigma^2 \int_{\frac{a-\mu}{\sigma}}^{\infty} z e^{-z^2/2} dz + \mu \int_{\frac{a-\mu}{\sigma}}^{\infty} e^{-z^2/2} \sigma dz = \sigma^2[-e^{-z^2/2}]_{\frac{a-\mu}{\sigma}}^{\infty} + \mu \int_a^{\infty} e^{-\frac{1}{2}\left(\frac{w-\mu}{\sigma}\right)^2} dw = \sigma^2 e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2} + \mu P\sigma\sqrt{2\pi}$$

Substituting (9) in (8),

$$(10)\quad \eta = -\frac{1}{P\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2} \cdot \left[\frac{\sigma^2}{P\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2} + \frac{\mu P\sigma\sqrt{2\pi}}{P\sigma\sqrt{2\pi}} - a\right] = -\frac{1}{P} f(a)\left[\frac{\sigma^2}{P} f(a) + \mu - a\right]$$

For any value of a, $f(a)$ and P can be calculated from standard statistical tables of the normal distribution. Then, η can be computed as shown in Table 1, and is seen to be a monotonically increasing function of P, always greater than -1.[5]

[3]Similar results are obtainable for the case of an infinite time horizon and a positive discount rate.

[4]While a general result could not be obtained for the *Gamma* distribution, two particular forms of this distribution were tried (one was a Chi-Square with four degrees of freedom) and the condition $\eta > -1$ was found to hold for both.

[5]This is true for P at least as small as .005; and lower values of P would clearly seem irrelevant to any real world search situation. $P = .005$ implies the searcher's reservation wage is set such that the average number of offers required before acceptance is 200 $(= 1/P)$.

TABLE 1—NORMAL DISTRIBUTION

P	a	η
.005	$\mu + 2.576\sigma$	$-.94$
.025	$\mu + 1.96\sigma$	$-.88$
.05	$\mu + 1.645\sigma$	$-.86$
.1	$\mu + 1.28\sigma$	$-.84$
.2	$\mu + .84\sigma$	$-.79$
.3	$\mu + .525\sigma$	$-.74$
.4	$\mu + .25\sigma$	$-.69$
.5	μ	$-.64$
.6	$\mu - .25\sigma$	$-.58$
.7	$\mu - .525\sigma$	$-.51$
.8	$\mu - .84\sigma$	$-.42$
.9	$\mu - 1.28\sigma$	$-.29$

Note: P = probability of accepting an offer; a = minimum acceptable offer; η = elasticity of P with respect to v (the number of vacancies).

Hence, under two alternative and often used assumptions[6] (normal and rectangular distributions), Barron's model implies $dD/dv < 0$ instead of the ambiguous theoretical result he presents. An increase in the number of vacancies v, increasing the probability of *receiving* a wage offer θ, will cause a less than proportionate decrease in the probability of accepting a wage offer P; the net effect will be to decrease the average duration of unemployment D. So, the empirical evidence he gives can be seen now as support for his theoretical model, rather than as merely suggestive of some unspecified relationship between vacancies and duration of unemployment.

[6]Note that the assumption of a symmetric wage offer distribution is more plausible in a model such as Barron's than in earlier search models in which a concentration of zero offers must be allowed to account for situations of no vacancies.

REFERENCES

J. M. **Barron**, "Search in the Labor Market and the Duration of Unemployment: Some Empirical Evidence," *Amer. Econ. Rev.*, Dec. 1975, *65*, 934–42.

J. **Johnston**, *Econometric Methods*, 2d ed., New York 1972.

NOTES

The ninetieth annual meeting of the American Economic Association will be held in New York City, December 28-30, 1977. The Employment Center will be open from December 27-30.

Annual Meeting Employment Center

The Employment Center at the 1977 annual meetings of the Allied Social Science Associations in New York City will begin operation on December 27, the day before sessions begin. Applicants and employers will be able to attend more sessions with a day set aside entirely for labor market transactions. This service will be located at the Convention Employment Center in the Americana Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5.00 P.M., December 28-29; and 9:00 A.M. to 12.00 noon, December 30.

Because the 1978 annual meeting comes at an early date in the academic year (August 29-31), the Executive Committee has decided to provide employment services at a later time. A job market will *not* be organized for the August meeting in Chicago The AEA will provide an organized market in December 1978 or January 1979 at a site yet to be selected. Only one placement service will be provided during the academic year 1978-79, and it will be separate from and later than the annual meeting.

Call for Papers for the 1978 Meetings

Members wishing to give papers or make suggestions for the program for the meetings to be held in Chicago, August 29-31, 1978, are invited to send their ideas to Professor Robert Solow, Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. Although most of the sessions sponsored by the American Economic Association will consist of invited papers, there will also be several sessions of noneconometric contributed papers (The sessions of contributed papers will not be published in the *Papers and Proceedings* issue to appear February 1979.) Proposals for invited sessions should be submitted as soon as possible. To be considered for the contributed sessions, abstracts of proposed (noneconometric) papers must be received no later than February 1, 1978. Economists wishing to give papers on econometrics or economic theory may submit abstracts to the Econometric Society, which meets with the American Economic Association and annually schedules a substantial number of contributions.

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings abroad that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Grants are likely to cover only lowest cost excursion fares and will rarely exceed 50 percent of full economy-class fares. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are ineligible. The deadlines for applications to be received in the office of the American Economic Association are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1, for meetings scheduled between March and June, November 1. Application forms may be obtained from C. Elton Hinshaw, Secretary, American Economic Association, 1313 21st Avenue South, Nashville, Tennessee 37212

For many years the American Economic Association has considered the computerization of its membership and subscription lists These lists are now on the computer beginning with this issue of the journals (December). In the future it will be extremely important that a copy of your invoice accompany your payments and a copy of your mailing label be sent with your changes of address For details, see the Advertising Section of this issue. If you have questions, please address them to the office of the Secretary of the Association.

The Joint Committee on Eastern Europe of the American Council of Learned Societies and the Social Science Research Council wishes to draw attention to its program of *Grants for Postdoctoral Research in East European Studies*. The deadline for receipt of application forms is December 30, 1977. Requests for information about this or other *ACLS* fellowships and grants programs should be addressed to Office of Fellowships and Grants, American Council of Learned Societies, 345 East 46 Street, New York, NY 10017.

The S.S. Huebner Foundation for Insurance Education, a nonprofit foundation affiliated with the Univer-

sity of Pennsylvania, is sponsoring its third annual research grants competition. The grants are intended to support research in the field of risk and insurance. Full-time faculty members at colleges and universities in the United States and Canada are eligible to apply for grants. Applicants should hold a terminal degree such as Ph.D. or D.B.A., a law degree, or be a fellow of an actuarial society. Among the general topic areas which will be considered for grants are insurance company investment strategies, risk management, risk theory, consumer demand for insurance, health insurance, other methods of health care financing, social insurance plans and programs, international insurance problems and issues, insurance law, and insurance regulation. Grants are available from the Foundation in amounts up to $10,000. Proposals must be submitted by March 1, 1978, and the grants will be awarded by May 1, 1978. Additional information regarding the program can be obtained by writing to Dr. J. David Cummins, Research Director, S.S. Huebner Foundation for Insurance Education, W-133 Dietrich Hall, University of Pennsylvania, Philadelphia, PA 19174.

For the academic year 1978-79 Harvard Law School offers four or five Liberal Arts Fellowships to college and university teachers in the arts and sciences for a year at the Harvard Law School. The purpose of the fellowships is to enable teachers in the social sciences or humanities to study fundamental techniques, concepts, and aims of law so that they will be better able to use legal materials and legal insights which are relevant to their own disciplines. Fellowship holders will presumably take at least two first-year courses in law, in addition to more advanced courses, and will participate in a joint seminar. The year of study will not count toward a degree. The fellowship grant covers tuition and health fees only. Applications should include a biographical resumé (including academic record and list of publications), a statement explaining what the applicant hopes to achieve through his/her year of study, and two letters of recommendation (mailed directly to the Chairman from the referees). There is no special application form. Applications for 1978-79 should be submitted before January 15, 1978, to the Chairman, Committee on Liberal Arts Fellowships, Harvard Law School, Cambridge, MA 02138. Awards will be announced before February 15, 1978.

The Conference on Social Sciences in Health, an affiliated organization of the American Public Health Association, is composed of individuals from the social and behavioral sciences and the health care field. The Conference provides an interdisciplinary national forum in which views of social and behavioral sciences are focussed on major questions in health and major issues of public policy in health. The Conference carries on its activities through scientific sessions and informal discussions at the annual meetings of the A.P.H.A. and a newsletter to its members. Information on, or application for, membership with annual dues of $5.00 is available from Ms. Anne Cugliani, Three Park Avenue, New York, NY 10016.

Mathematica, Inc. invites applications for the Oskar Morgenstern Distinguished Fellowship, at Mathematica. The purpose of the fellowship is to enable a member of the academic or research staff of a university, an official of the *U.S.* government, or a researcher elsewhere, to spend a sabbatical leave at Mathematica. The fellow will be able to continue personal research, participate in Mathematica's technical and scientific activities, review Mathematica's project reports or technical work in his/her field, and will present two lectures, the "Oskar Morgenstern Lectures," before Mathematica's staff and guests. Recommendations, applications, and inquiries should be sent to Dr. Tibor Fabian, President, Mathematica, P.O. Box 2392, Princeton, NJ 08540 by October 15 of the academic year preceding the year of application.

The Exxon Education Foundation has funds available under its Economics and Financing of Higher Education Program to underwrite both research and pilot projects designed to develop an understanding of the full economic, social, and political consequences of proposed methods of financing higher education. The program's aim is to stimulate the analysis by funding pilot demonstrations of untried methods of financial support. The approaches used to achieve these objectives are grants to teams of competent social scientists to investigate, analyze, and compare the effects of funding procedures previously used or currently in use, and larger grants to consortia, educational organizations, and public and private agencies to implement and study small scale applications of funding systems that are as yet untested. For information, write Exxon Education Foundation, 111 West 49th Street, New York, NY 10020.

New Journal

The Indian Association for Research in Income and Wealth will publish *The Journal of Income and Wealth.* It will appear twice yearly in April and October, and be edited by Professor M. Mukherjee. For additional information, contact Mahinder D. Chaudhry, Department of Political and Economic Science, Royal Military College of Canada, Kingston, Ontario K7L 2W3.

The Eastern Community College Social Science Association will hold its 1978 conference at Grossinger's Conference Center in Grossinger, New York, April 2-4, 1978. The Program Chairman is Arnold Toback, Social Science Department, Dutchess Community College, Pendell Road, Poughkeepsie, NY 12601.

The Center for Defense Information announces a Junior Fellowship Program for advanced graduates who intend to devote a substantial part of their early professional careers to policy-related research. The program seeks to bridge formal academic preparation and project experience by associating junior fellows closely with the Center's staff and research objectives. For this reason, junior fellows will participate full time in Center activities and undertake projects compatible with the scope of its interests and publications. Appointments should not be sought to pursue academic or degree requirements. Prospective junior fellows should have a recent Ph.D. or be reasonably close to completing degree requirements. Individuals at the M.A. level will be considered under exceptional circumstances, such as unusual achievements or prior experience in relevant professional work. Applicants should have strong backgrounds in one or more of the following areas: military studies, policy analysis, political science, public administration, law. For information, contact Dr. Robert M. Whitaker, Staff Director, Center for Defense Information, 122 Maryland Ave., NE, Washington, D.C. 20002.

A newsletter to be called *Lex/econ* serving the law and economics discipline has been announced by the Law and Economics Center of the University of Miami School of Law. The first issue is scheduled for December 1977. *Lex/econ* will carry announcements of forthcoming conferences, seminars, institutes, and other programs of interest to law and economics scholars, available manuscripts in circulation for criticism or already accepted for publication (but not yet out), and job openings and appointments in the field. The newsletter will also seek to call attention to speeches, papers, lectures, new books, and articles of interest in specialized publications. When possible it will report the availability of bibliographies Letters and news announcements are welcomed. Address Editor, *Lex/econ*, Law and Economics Center, P.O Box 248000, Coral Gables, FL 33124

Omicron Delta Epsilon, the International Honor Society in Economics, invites the submission of entries for the tenth year of the Irving Fisher Graduate Monograph and Frank W. Taussig Award Undergraduate Competitions. The Fisher Award consists of $1,000 and publication as a book by Princeton University Press, subject to approval of its editorial board. In addition, the winner will be invited to submit a paper based on the winning entry to the *American Economic Review*. The recommendations of the Final Selection Board of the Competition will be considered by the *Review* in the refereeing process. All finalists will be invited to submit a paper for publication in *The American Economist*. The Taussig Award consists of $100 and publication in *The American Economist* Entries for the Fisher Award should be submitted to Departmental Selection Committees by January 1, 1978, and entries for the Taussig Award by May 15, 1978. They will be judged by the International Editorial Board and finalists by the Final Selection Board, consisting of Professors Frank H. Hahn, Dale W. Jorgenson, Robert M. Solow, Arnold Zellner, and Egon Neuberger (editor). Anyone interested in entering the competitions should contact Egon Neuberger, Editor, Economic Research Bureau, State University of New York, Stony Brook, NY 11794.

The Inter-University Consortium for Political and Social Research (*ICPSR*) is designed to facilitate research and instruction in the social sciences. The *ICPSR* provides. a central repository and dissemination service for computer-readable social science data; training facilities in basic and advance methods of quantitative social analysis; and resources for facilitating use of advanced computer technology The *ICPSR* is a partnership between over 200 member academic institutions in the United States, Canada, and other nations, and the Center for Political Studies of the Institute for Social Research, University of Michigan.

The data repository receives, processes, and distributes computer-readable data relevant to over 130 countries. The archive now includes almost 500 data collections. A comprehensive guide to data holdings and services is available on request. Examples of current holdings of potential interest to economists include data on such topics as family income, surveys of consumer attitudes, behavior and finances, working conditions, technology and labor, and revenue sharing in addition to census, population, and income data for the United States and other countries. Students and faculty at member institutions have access to data without charge beyond the institutional membership fee.

The *ICPSR* aids its members in efforts to eliminate barriers to utilization of computer technology. The staff provides information on developments in computer hardware, software, and data management and analysis packages. *ICPSR* disseminates an integrated package of computer programs (*OSIRIS*) which provides extensive data management and analysis capabilities including tabulation routines, correlation and regression techniques, and more advanced multivariate, nonparametric, and dimensional analysis procedures Inquiries about *ICPSR*, its data holdings, and services should be addressed to Executive Director, *ICPSR*, P.O. Box 1248, Ann Arbor, MI 48106.

A World Congress will be held in Córdoba, Argentina, June 1978, devoted to the analysis of law, sociology, medicine, and economics of sports. It is part of the activities attendant on the World Soccer Cup An economics section will be devoted to two topics National Accounts—Contributions regarding the structure and participation of sport within the gross domestic product (methodological and empirical papers); and Economic Analysis—Economic theory and sports: Microeconomic aspects (sports clubs' be-

havior); the football (or any other sport) player; market structure and sport organization; and so on; Macroeconomic aspects consisting of analysis of the impact of sport activities on the economy (economic and sociopolitical and cultural effects); and empirical and comparative studies (among different activities and/or countries, or through time); the football (soccer) phenomenon, its economic implication. For further information, write to Professor Luis Eugenio Di Marco, University of Córdoba, Argentina.

Call for Papers

The sixth annual Telecommunications Policy Research Conference will be held at Airlie House, Virginia, May 10-13, 1978. The conference brings public and private policy makers together with researchers in all areas of telecommunications policy. The conference organizing committee is soliciting *outlines of research papers* that bear, directly or indirectly, on questions in these areas. Some of the submitted papers will be chosen for presentation at the conference on the basis of the quality and relevance of the research. Travel and conference living expenses will be reimbursed if no alternative source of funding is available. Please send abstracts, name, address, telephone number, and professional affiliations *immediately* to: Telecommunications Policy Research Conference Organizing Committee; c/o R. D. Willig, Bell Laboratories, Holmdel, NJ 07733

The department of economics of the City College of the City University of New York will hold its annual conference on May 11, 1978, on the subject, "Law and the Economy." Papers are invited on any aspect of the interaction of law and economics. Inquiries should be addressed to Professor Gerald Sirkin, Department of Economics, City College, City University of New York, Convent Avenue and 138th Street, New York, NY 10031. Papers delivered at the conference will be published in the sixth conference volume

---

Deaths

Arthur Ford, associate professor of economics, Southern Illinois University, June 16, 1977.

Jacob Marschak, University of California-Los Angeles, July 27, 1977

Oskar Morgenstern, chairman of the Board of Directors of Mathematica, Inc., July 26, 1977.

Jim E. Reese, professor of economics, University of Oklahoma, Oct. 6, 1976.

Jack L. Robinson, professor of economics, University of Oklahoma, Dec. 7, 1975.

Retirements

Julian H. Bradsher, professor of economics, Oklahoma State University, June 1977.

Harold W. Davey, professor of economics, Iowa State University, May 31, 1977.

Dwight P. Flanders, professor of economics, University of Illinois, Aug. 1977.

Charles P. Kindleberger, professor of economics emeritus and senior lecturer, Massachusetts Institute of Technology, July 1, 1976.

Will E. Mason, professor of economics emeritus, Pennsylvania State University, June 30, 1977.

Charles L. Merwin, deputy director, African Department, International Monetary Fund, July 1977.

Ralph B. Price, professor of economics, Western Maryland College, June 1977.

John Simpson, department of economics, Eastern Michigan University, June 1977.

Randall S. Stout, professor of economics emeritus, Pennsylvania State University, June 30, 1977.

Visiting Foreign Scholars

Ernst Berndt, University of British Columbia: visiting scholar, Massachusetts Institute of Technology, Sept. 1, 1977.

Tuvia Blumenthal, Australian National University: visiting professor of economics, University of Illinois, Jan. 1978.

Beat Burgenmeier, University of Geneva: postdoctoral fellow, Massachusetts Institute of Technology, Sept. 1, 1977.

Ting-An Chen, Munster University, West Germany: visiting scholar, Massachusetts Institute of Technology, Sept. 1, 1977.

Peter A. Cornelisse, Erasmus University, Rotterdam: visiting associate professor, department of economics, Cornell University.

Elhanan Helpman, University of Tel-Aviv: visiting associate professor, department of economics, University of Rochester, July 1, 1977.

Vesa Kanniainen, Helsinki University: visiting assistant professor of economics, Brown University, 1977-78.

Shlomo Maital, Tel-Aviv University: visiting lecturer, Woodrow Wilson School of Public and International Affairs, Princeton University, 1977-78.

Grayham E. Mizon, London School of Economics and Political Science: visiting foreign scholar, University of California-San Diego, fall 1977

Seiichi Ota, Fukuoka University, Japan: visiting assistant professor of economics, Brown University, 1977-78.

Poul Rasmussen, University of Copenhagen: visiting professor of economics, University of Illinois, Jan. 1978.

Elke Schaefer, University of Saarbrucken, West Germany: visiting scholar, Massachusetts Institute of Technology, Sept. 1, 1977.

Avia Spivak, Ben Gurion University of the Negev, Israel: visiting assistant professor of economics, Brown University, 1977-78.

Arne Dag Sti, Norwegian School of Economics and Business Administration: visiting scholar, Massachusetts Institute of Technology, Aug. 1, 1977.

Menahem Yaari, Hebrew University: research associate, Massachusetts Institute of Technology, July 1, 1977.

Promotions

Jerald R. Barnard: professor, department of economics, University of Iowa, Aug. 1977.

Larry G. Beall: associate professor of economics, Virginia Commonwealth University, Sept. 1977.

Ashok Bhargava: associate professor of economics, University of Wisconsin-Whitewater, July 1, 1977.

Michael D. Boehlje: professor of economics, Iowa State University, July 1, 1977.

Stephen V. O. Clarke: senior economist, research and statistics function, Federal Reserve Bank of New York, June 1, 1977.

Warren T. Dent: professor, department of economics, University of Iowa, Aug. 1977.

Arnold M. Faden: professor of economics, Iowa State University, Sept. 1, 1977.

Allan M. Feldman: associate professor, department of economics, Brown University, July 1, 1977.

Stanley Fischer: professor of economics, Massachusetts Institute of Technology, July 1, 1977.

Margaret L. Greene: vice president, foreign function, Federal Reserve Bank of New York, June 1, 1977.

Ann K. Harper associate professor of economics, Western Maryland College, Sept. 1, 1977.

Paul L. Joskow. associate professor, Massachusetts Institute of Technology, July 1, 1977.

Richard E. Kihlstrom: professor of economics, University of Illinois, Aug. 1977.

Jene K. Kloon· professor of economics, department of economics, Northern Illinois University, Aug. 1977.

Ramon Knauerhase: professor, department of economics, University of Connecticut, Oct. 1977.

Woo Bong Lee: associate professor, department of economics, Bloomsburg State College.

Mukul Majumdar: professor, department of economics, Cornell University, Feb. 1976.

Richard C Maxon· professor of economics, Iowa State University, July 1, 1977

Robert P. Parks: associate professor of economics, Washington University, July 1977.

Andrew Postlewaite. associate professor of economics, University of Illinois, Aug. 1977.

Ronald E. Raikes· associate professor of economics, Iowa State University, July 1, 1977.

Raymond Riezman· assistant professor, department of economics, University of Iowa, Jan 1977.

Anthony A. Romeo: associate professor, department of economics, University of Connecticut, Oct. 1977.

Richard Rosenberg: associate professor of economics, Pennsylvania State University, July 1, 1977.

Stephen R. Sacks: associate professor, department of economics, University of Connecticut, Oct. 1977.

Takamitsu Sawa: professor of economics, University of Illinois, Aug. 1977.

John B. Shoven: associate professor of economics, Stanford University, Sept. 1977.

J. Kirker Stephens: professor of economics, University of Oklahoma, Sept. 1, 1977.

William C. Wheaton: associate professor of economics and urban studies, Massachusetts Institute of Technology, July 1, 1977.

Robert N. Wisner: professor of economics, Iowa State University, July 1, 1977.

Administrative Appointments

Morton S. Baratz: General Secretary of the American Association of University Professors, June 14, 1977.

Henry N. Goldstein: professor of economics, department head, University of Oregon, fall 1977.

James E. Jonish: chairman, department of economics, Texas Tech University, Sept. 1, 1977.

Martin T. Katzman, Harvard University: head, graduate program in political economy, University of Texas-Dallas, Sept. 1977.

Gerald M. Lage. interim chairman, department of economics, Oklahoma State University, May 15, 1977.

Alton D Law· chairman, department of economics, Western Maryland College, Sept. 1977.

Woo Bong Lee: chairman, department of economics, Bloomsburg State College, June 1, 1977.

Norman Mintz: deputy provost, Columbia University, Aug. 1, 1977.

J. Eugène Poirier. chairman, department of economics, Georgetown University, July 1, 1977

Robert W. Resek. director, bureau of business and economic research, University of Illinois, Aug. 1977.

Jules J. Schwartz, University of Pennsylvania. dean, School of Management, Boston University.

Sheldon W Stahl, Federal Reserve Bank of Kansas City: deputy governor, office of finance and research, Farm Credit Administration, Washington, D.C , Aug 1977.

Stanley W. Steinkamp. vice chairman of economics, University of Illinois, Aug. 1977.

J Kirker Stephens· division director, department of economics, University of Oklahoma, Oct. 19, 1976

Bernard L. Weinstein. director, Southwest Center for Economic and Community Development, University of Texas-Dallas, July 1977.

Appointments

Roger Alcaly: economist, Banking Studies Division, Federal Reserve Bank of New York, Aug. 9, 1977.

Susan Alexander: assistant professor, department of economics, University of Iowa, fall 1977.

William T. Alpert: assistant professor of economics, Washington University, Sept. 1977.

John Anderson, Claremont Graduate School: assistant professor, economics department, Eastern Michigan University, Sept. 1977.

Robert Averitt, Smith College: visiting professor of political economy, University of Texas-Dallas, Sept. 1977.

Randall S. Bausor, Duke University: assistant professor of economics, Virginia Commonwealth University, Aug. 1977

Katherine C. Bazan: visiting assistant professor of economics, College of William and Mary, 1977–78.

Gordon Bennett, Canadian Ministry of Natural Resources: assistant professor, economics department, Eastern Michigan University, Sept. 1977.

Charles Bischoff, IBM: associate professor of economics, State University of New York-Binghamton, Sept. 1, 1977.

Leigh B. Boske, Wisconsin Department of Transportation: senior economist, National Transportation Policy Study Commission, June 1977.

Samuel H. Bostaph, Hamilton College: assistant professor, department of economics, Western Maryland College, Sept. 1, 1977.

Richard Boyce, Harvard University. assistant professor of economics, State University of New York-Binghamton, Sept. 1, 1977.

J. A. Cacy: vice president and senior economist, Federal Reserve Bank of Kansas City, Aug. 1, 1977.

W. Michael Cox, Virginia Polytechnic Institute and State University: assistant professor, department of economics, University of Rochester, Sept. 1, 1977.

Steven M. Crafton, University of Tennessee. assistant professor of economics, Virginia Commonwealth University, Aug. 1977.

Lesley D Daniels assistant professor of economics and urban studies, Washington University, Sept 1977.

Robert C. DauffenBach, University of Illinois-Urbana assistant professor of economics, Oklahoma State University, Sept. 1, 1977.

Tulinda Deegan, U.S Department of Transportation. assistant economist, National Transportation Policy Study Commission, July 1977

Edwin G. Dolan, Dartmouth College· staff economist, Economic Policy Office, Antitrust Division, U S. Department of Justice, Oct 1, 1977

Randall Eberts, Northwestern University assistant professor, department of economics, University of Oregon, Sept. 1977.

Liam P. Ebrill, Harvard University· assistant professor, department of economics, Cornell University.

Margo B. Faier, University of Pennsylvania. staff economist, Economic Policy Office, Antitrust Division, U S. Department of Justice, Sept. 1977

Ray Fair, Yale University· visiting associate professor of economics, Massachusetts Institute of Technology, fall 1977.

Henry S. Farber: assistant professor of economics, Massachusetts Institute of Technology, Sept 1, 1977.

John H. Gates visiting assistant professor of economics, College of William and Mary, 1977–78.

Mark L. Gertler, Stanford University: assistant professor, department of economics, Cornell University.

Ronald D. Gilbert, Louisiana State University. associate professor, department of economics, Texas Tech University, Sept 1, 1977.

John C. Goodman, Dartmouth College· visiting assistant professor of economics, Southern Methodist University, Sept. 1977.

John Graham, Northwestern University· lecturer in economics, University of Illinois, Aug. 1977

Earl L. Grinols, Massachusetts Institute of Technology: assistant professor, department of economics, Cornell University.

Jacob Grossman: economist, Market Statistics Division, Federal Reserve Bank of New York, May 31, 1977.

Joseph Halevi: assistant professor, department of economics, Rutgers-The State University, July 1, 1977.

Raouf Hanna, Colby College: assistant professor, economics department, Eastern Michigan University, June 1977.

Donald A. Hanson, Massachusetts Institute of Technology: assistant professor of economics, Southern Methodist University, Sept 1977

Ronald Harstad, University of Pennsylvania: lecturer in economics, University of Illinois, Aug. 1977.

George A. Hay, U.S Department of Justice· visiting professor, department of economics and the Law School, Cornell University.

Steven Hayworth, Massachusetts Institute for Technology· assistant professor, Eastern Michigan University, Sept 1977

Roger T. Kaufman, Massachusetts Institute of Technology assistant professor, Amherst College, July 1, 1977.

Harry C Katz assistant professor of economics and management, Massachusetts Institute of Technology, Sept. 1, 1977

Douglass J Klein, Syracuse University Brookings fellow, Washington, D.C., July 1977

David R. Knowles, Washington State University: staff economist, Economics Policy Office, Antitrust Division, U S Department of Justice, Oct 1, 1977.

Myron L. Kwast assistant professor of economics, University of Oklahoma, Sept. 1, 1977

Lucinda M Lewis, University of Wisconsin staff economist, Economic Policy Office, Antitrust Division, U.S Department of Justice, Oct. 1, 1977

David M Lilien, Massachusetts Institute of Technology assistant professor, department of economics, University of California-San Diego, July 1, 1977

Peter H. Lindert, University of Wisconsin· professor of economics, University of California-Davis, July 1977.

Raymond E Lombra, Federal Reserve System associate professor of economics, Pennsylvania State University, Apr. 1, 1977.

Thomas S McCaleb, University of Kansas: visiting assistant professor, Rice University, July 1977-July 1979

John H. McDermott, Brown University. assistant professor of economics, College of the Holy Cross, Sept. 1, 1977.

Henry McFarland, Northwestern University staff economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, Oct. 1, 1977

Brian McGrath, Brown University· assistant professor of economics, Indiana University, Sept. 1, 1977.

Paul D. McNelis, Weston School of Theology: assistant professor, department of economics, Georgetown University, Aug. 29, 1977.

Richard MacMinn, University of Illinois: assistant

professor of economics, State University of New York-Binghamton, Sept. 1, 1977.

Thomas MaCurdy, University of Chicago: assistant professor, department of economics, Stanford University, Sept. 1977.

Eric S. Maskin: assistant professor of economics, Massachusetts Institute of Technology, Sept. 1, 1977.

Y.P. Mehra, Illinois State University: professor of economics, Pennsylvania State University, Sept. 1, 1977.

Leonard Merewitz, J. W. Wilson & Associates, Inc.: senior economist, National Transportation Policy Study Commission, July 1977.

Charles L. Merwin, International Monetary Fund: adjunct professor, department of economics, American University, Aug. 1977.

Hajime Miyazaki, University of California-Berkeley: assistant professor, department of economics, Stanford University, Sept 1977.

Wilhelm Neuefeind, Universität Bonn. professor of economics, Washington University, Jan. 1978.

Richard S. Newfarmer, University of Wisconsin-Madison: assistant professor, University of Notre Dame, Sept. 1977

Prasanta K. Pattanaik, La Trobe University, Australia: professor of economics, Southern Methodist University, Sept. 1977.

Joseph D. Reid, Jr., University of Chicago: associate professor of economics, Southern Methodist University, Sept. 1977.

Helen Reynolds, Southern Methodist University: assistant professor, University of Texas-Dallas, Sept. 1977

Edward Rice, University of California-Los Angeles: lecturer in economics, University of Illinois, Aug 1977.

Kevin W. S. Roberts assistant professor of economics, Massachusetts Institute of Technology, Sept. 1, 1977.

Vernon Vance Roley: financial economist, Federal Reserve Bank of Kansas City, June 1977

Marilyn J. Simon. assistant professor of economics, Massachusetts Institute of Technology, Sept. 1, 1977.

Dickson K Smith, Marquette University. associate professor of economics, Lakeland College, Aug. 29, 1977.

James L. Smith, Harvard University. lecturer in economics, University of Illinois, Aug 1977.

Larry R. Spancake, Yale University. instructor, department of economics, Fordham University, Sept. 1977.

David E. Spencer, Illinois State University: visiting assistant professor of economics, Arizona State University, Aug. 22, 1977.

Rita C. Sweeney, Princeton University: instructor, department of economics, Fordham University, Sept. 1977.

Thomas F. Tabasz, Miami University: associate professor of economics, College of Business and Economics, Washington State College.

Samuel Taddesse: economist, Banking Studies Division, Federal Reserve Bank of New York, May 31, 1977.

Richard K. Taube, Wisconsin Department of Transportation: director of policy development, National Transportation Policy Study Commission, June 1977.

William E. Taylor, Bell Laboratories: visiting associate professor of economics, Massachusetts Institute of Technology, Sept. 1-Dec. 31, 1977.

Mai-Trang Tran, Southern Illinois University: assistant professor, department of economics Texas Tech University, Sept. 1, 1977.

Thomas Ulen, Stanford University: lecturer in economics, University of Illinois, Aug. 1977.

George M. Vredeveld, University of Missouri: associate professor of economics and director, Greater Cincinnati Center for Economic Education, University of Cincinnati, July 1977.

Michael J. Wasylenko, University of Wisconsin: assistant professor of economics, Pennsylvania State University, Sept. 1, 1977.

Barry R. Weingast, Washington University: assistant professor of economics and research associate, Center for the Study of American Business, Sept. 1977.

Nancy Wentzler, University of Wisconsin: assistant professor of economics, Pennsylvania State University, Sept 1, 1977.

Gregory J. Werden, University of Wisconsin: staff economist, Economic Policy Office, Antitrust Division, U S. Department of Justice, Oct. 1, 1977.

Eden S H. Yu. assistant professor of economics, University of Oklahoma, Sept 1, 1977

Itzhak Zilcha, University of Illinois: assistant professor, department of economics, Cornell University.

Leaves for Special Appointments

Richard J. Arnould, University of Illinois: visiting research associate professor, Duke University, 1977-78.

William A. Brock, Cornell University: department of economics, University of Wisconsin.

Thomas S. Friedland, University of Illinois: Harper, Inc., New York City, 1977-78

John W. Fuller, Wisconsin Department of Transportation deputy director, National Transportation Policy Study Commission, 1977-78

David Greytak, Syracuse University: Center for Urban Studies, Johns Hopkins University, Sept. 1977.

Clyde A. Haulman, College of William and Mary visiting associate professor, department of economics and finance, Florida Technological University, 1977-78.

James M. Holmes, State University of New York-Buffalo: visiting research professor, Arizona State University, Jan. 20, 1978.

John W. Hooper, University of California-San Diego: School of Industrial Management, University of Petroleum and Minerals, Dhahran, Saudi Arabia, 1977-79.

George G. Judge, University of Illinois: professor of economics, University of Georgia, 1977-78.

Alfred E. Kahn, Cornell University: chairman, Civil Aeronautics Board, Washington, D.C.

Cotton Mather Lindsay, University of California-Los Angeles: visiting research professor, Arizona State University, Aug. 22, 1977.

Charles E. McLure, Jr, Rice University: executive director for research, National Bureau of Economic Research, Boston, June 1977–June 1979.

Masahiro Okuno, University of Illinois: visiting research assistant professor, Kyoto University, Japan, 1977–78.

Ibrahim M. Oweiss, Georgetown University: supervisor, Egyptian Economics Office, Ministry of Economy and Economic Cooperation, New York, 1977–78.

David Puryear, Syracuse University: Department of Housing and Urban Development, Washington, D.C., Sept. 1, 1977.

Richard E. Schuler, Cornell University. director, Office of Research, New York State Public Service Commission.

Roger B. Skurski, University of Notre Dame: Centre for Russian and East European Studies, University of Birminham, England, 1977–78.

Steven Slutsky, Cornell University: research fellowship, CORE, Universite Catholique de Louvain, Belgium.

Thomas Sowell, Center for Advanced Study in the Behavioral Sciences: visiting professor, Amherst College, July 1977.

Resignations

Hugh Folk, University of Illinois: University of Hawaii, Aug. 1977.

William A. Hyman, Wisconsin Department of Transportation, Sept. 1, 1977.

Roger Koenker, University of Illinois: Bell Laboratories.

Dale J. Poirier, University of Illinois, Aug. 1977.

Wolfhard Ramm, University of California-San Diego: Board of Governors, Federal Reserve Board, Aug. 1, 1977.

Sherwin Rosen, University of Rochester: University of Chicago, July 1, 1977.

Richard L. Schmalensee, University of California-San Diego. Sloan Institute, July 1, 1977.

Hal R. Varian, Massachusetts Institute of Technology, June 31, 1977.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A Please use the following categories:

1–Deaths
2–Retirements
3–Foreign Scholars (visiting the USA or Canada)
4–Promotions
5–Administrative Appointments
6–New Appointments
7–Leaves for Special Appointments (NOT Sabbaticals)
8–Resignations
9–Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), and the date at which the change will occur.

C Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

SEVENTY-FOURTH LIST OF DOCTORAL DISSERTATIONS IN POLITICAL ECONOMY IN AMERICAN UNIVERSITIES AND COLLEGES

The present list specifies doctoral degrees conferred during the academic year terminating June 1977. Abstracts will no longer be printed, as they are published by University Microfilms, Ann Arbor, Michigan.

General Economics; including Economic Theory, History of Thought, Methodology, Economic History, and Economic Systems

WILLIAM G. BENTLEY, Ph.D. Georgia State 1977 Wealth distribution in colonial South Carolina.

DOUGLAS H. BLAIR, Ph.D. Yale 1976. Essays in social choice theory.

DURCARMEL BOCAGE, Ph.D. Catholic 1976. The economic domination theory of Francois Perroux.

YUN HWANG BOO, Ph.D. Oklahoma 1977. An empirical comparison of theoretical models for a utility under separability hypotheses. A system of derived demand equations of fuels for *U.S* household use, 1937 70.

AVISHAY BRAVERMAN, Ph.D Stanford 1976. Monopolistic competition due to consumers' imperfect information.

FLINT BRAYTON, Ph.D. Johns Hopkins 1977 Rational inflation expectations A macroeconomic model and empirical tests

THOMAS P. BRESLIN, Ph.D West Virginia 1976. Bituminous coal· An export base for West Virginia, 1870 1930.

GREGORY BULLEN, Ph.D. Oregon 1976. Consumer saving through asset adjustment· An analysis of theoretical implications and test of an aggregate model.

KENNETH BURDETT, Ph.D Northwestern 1976 Essays on markets with imperfect price information.

LOUIS J. CHERENE, Ph.D. Wisconsin (Madison) 1976. Set valued dynamical systems and economic theory

PAUL P. CHRISTENSEN, Ph D. Wisconsin (Madison) 1976 Land, labor, and mechanization of the antebellum *U S.* economy

JAMES E. CLARK, Ph.D. Northwestern 1977. The impact of transportation technology on suburbanization in the Chicago region, 1830-1920.

SUSAN I. COHEN, Ph.D. Northwestern 1977. Incentive compatible control of the multidivisional firm with iterative communication.

RAYMOND COHN, Ph.D. Oregon 1977. A locational analysis of manufacturing activity in the antebellum South and Midwest.

CHARLES E. CONROD, Ph.D. Northwestern 1976 Limited growth of cities in the lower Ohio valley.

PETER J. COUGHLIN, Ph.D. State University of New York (Albany) 1976. A theory of disequilibrium decision making under uncertainty with an application to labor markets.

ROBERT A. DICKLER, Ph.D. Pennsylvania 1975. Labor market pressure and agricultural growth in the eastern region of Prussia, 1840-1914.

DONALD R DOHRMANN, Ph.D. Yale 1976 Screw propulsion in American lake and coastal steam navigation, 1840-60.

JAMES DOTI, Ph.D. Chicago 1976. The response of economics to environmental influences.

ROLF H. DUMKE, Ph.D. Wisconsin (Madison) 1976. The political economy of German economic unification: Tariffs, trade, and politics of the Zollverein era.

JONATHAN EATON, Ph.D. Yale 1976. Four essays in the theory of uncertainty and portfolio choice.

LARRY G EPSTEIN, Ph.D. British Columbia 1977 Essays in the economics of uncertainty.

LINWOOD T. GEIGER, Ph.D. Temple 1977. An economic analysis of expropriation.

JOHN L. GIBBENS, Ph.D. Virginia Polytechnic Institute 1977. William E. Hearn and classical political economy.

JOHN GODFREY, Ph.D. Georgia 1976. Monetary expansion in the Confederacy.

JOSE L. GUASCH-A, Ph.D. Stanford 1976 Anticipations and diversity in intemporal economies

THORVALDUR GYLFASON, Ph D. Princeton 1977. Inflation, unemployment, and economic growth· Two essays

PETER K HAMMERSCHMIDT, Ph.D Colorado State 1976 Community A nonutilitarian approach to economics.

HARRISON HARTMAN, Ph.D New York 1976. The theoretical debate on income policy in the United States since 1960· A critique

JAMES HENDERSON, Ph.D. Northern Illinois 1977 Early British economics, 1822 50

IAN R.C. HIRST, Ph.D. Chicago 1976. Efficiency in trading and the revelation of preferences.

CHARLES A. HOLT, JR., Ph.D. Carnegie-Mellon 1977 Bidding for contracts.

HAMID HOSSEINI, Ph.D Oregon 1977. Ricardo's theory of comparative costs reconsidered.

THOMAS HUERTAS, Ph.D. Chicago 1977. Economic growth and economic policy in a multinational setting. The Hapsburg Monarchy, 1841-65.

WATANA ISARANKURA, Ph.D. California (Santa Barbara) 1977. Determinants of the income distribution in the United States: A macroeconomic approach.

GERALD D. JAYNES, Ph.D. Illinois 1976. Essays in the economics of imperfect information.

WARREN J. JESTIN, Ph.D. Toronto 1977. Provincial

policy and the development of the metallic mining industry in Northern Ontario: 1845–1920.

DANA B. JOHNSON, Ph.D. Northwestern 1977. A study of the relationship of production decisions to finished inventory and unfilled orders.

JAMES R. JOHNSON III, Ph.D. Duke 1977. Toward a characteristic space theory of impure public goods.

AKIO KAGAWA, Ph.D. Rochester 1977. Essays on equilibrium and disequilibrium.

JOSEPH E. KECKEISSEN, Ph.D. New York 1976. The meanings of economic law.

WILLIAM A. KELLY, JR., Ph.D. North Carolina (Chapel Hill) 1976. The theory of price and quantity adjustments in an N-firm market.

ROGER C. KORMENDI, Ph.D. California (Los Angeles) 1977. On the nature of decentralized pure exchange.

LARRY L. LAWSON, Ph.D. Colorado 1976. The concept of the individual economic theory.

FRANK D. LEWIS, Ph.D. Rochester 1977. Explaining the shift of labor from agriculture to industry in the United States: 1869–99.

LUCINDA M. LEWIS, Ph.D Yale 1977. Essays on purely competitive intertemporal exchange.

ALBERT N. LINK, Ph.D. Tulane 1976. The microfoundations of technological change. A theoretical and empirical model

STEPHEN A McCAFFERTY, Ph.D. Brown 1977. A theory of search for transaction partners

GARY M MARTIN, Ph.D. North Carolina (Chapel Hill) 1976. Explaining economic development in the southern United States, 1880 1930: An evaluation of neoclassical and neomercantilist models as they apply to a backward or developing region.

NICOLAS J MATHIEU, Ph D. Pennsylvania 1976 An application of control theory to macroeconomic policies in the Canadian economy

WALTER S. MISIOLEK, Ph.D. Cornell 1976. Price expectations, the real rate of interest, and the Phillips curve

JOHN P MONDEJAR, Ph D Indiana 1976. Neocolonialism as an economic system: Cuba, 1898–1934.

TAKESHI MUROTA, Ph.D. Minnesota 1976. Public information and social welfare under five alternative market mechanisms.

DICK K NANTO, Ph.D. Harvard 1977. The United States' role in the postwar economic recovery of Japan

PAMELA J. NICKLESS, Ph.D. Purdue 1976. Changing labor productivity and the utilization of native woman workers in the American cotton textile industry: 1825–60.

OSAMU NISHIMURA, Ph.D. Pennsylvania 1976. A dynamic theory of income distribution over the life cycle.

CARLOS F.D. OBREGON, Ph.D. Colorado 1976. Economics and the inquiry into social harmony.

MIKE J. O'BRIEN, D. Sci. Washington (St. Louis) 1976. Approachability of a macroeconomic model.

HERAKLIS M. POLEMARCHAKIS, Ph.D. Harvard 1977. Intertemporal allocations and the fixed price method.

ARTHUR R. PREISS, Ph.D. Indiana 1977. A microeconomic approach to business investment: Adjustment costs and the theory of the firm.

SALIM RASHID, Ph.D. Yale 1976. Economies with infinitely many traders.

JOSEPH V. REMENYI, Ph.D. Duke 1976. Core-demicore interaction in economics.

FRANK ROOSEVELT, Ph.D. New School 1977. Towards a Marxist critique of the Cambridge school.

ROY ROTHEIM, Ph.D. Rutgers 1977. Foundation of nonequilibrium economic analysis.

BERNARD ROTHMAN, Ph.D. Iowa State 1977. Explorations in the theory of the multiproduct firm.

RICHARD P. ROZEK Ph.D. Iowa 1976. A nontatonnement bidding for a centralized market.

ROBERT SANDY, Ph.D Michigan State 1977. Two essays in the history of short-run labor supply theory.

JOSEPH E. SANTANGELO, Ph.D. Cincinnati 1976. Nominal rates and price expectation in a complete income determination model.

SCOTT SAUNDERS, Ph.D. Johns Hopkins 1977. The existence of equilibrium in a second best model.

GITA SEN, Ph.D. Stanford 1976 The theory and estimation of disequilibrium markets.

BRYAN E. STANHOUSE, Ph.D. Illinois 1976 Stochastic control experiments on a small macro model under two expectations regimes.

JOHN N. STEVENS, Ph.D. Pennsylvania State 1977. The political economy of Czechoslovak foreign trade.

STEPHEN H. STRAND, Ph.D. Vanderbilt 1976. An analysis of decreasing costs in the production and transportation of a single product in spatially separated markets.

GERRY L. SUCHANEK, Ph.D. Northwestern 1977. Information and optimality in pollution control institutions.

KEN-ICHI TATSUMI, Ph.D. Pennsylvania 1975. Essays on the theory of the firm and financial market.

GEORGE TAVLAS, Ph.D New York 1977. Essays on the doctrinal historical development of Friedman's monetary economics.

LOUIS W. THOMSON, Ph.D. Stanford 1976. Incentives and information.

ROSS D. THOMSON, Ph.D. Yale 1976. The origin of modern industry in the United States: The mechanization of shoe and sewing machine production.

PETER G. TOUMANOFF, Ph.D. Washington 1977. The effects of property rights on resource allocation in Russia, 1861–1913.

BRUCE VAVRICHEK, Ph.D. Northwestern 1977 Consumer decision process with incomplete information.

SUSAN B. VROMAN, Ph.D. Johns Hopkins 1977. A theory of money wage dynamics with a heterogeneous labor force.

GORDON E. WAGNER, Ph.D. Cornell 1977 Consecration and stewardship: A socially efficient system of justice.

RONALD S. WARREN, JR., Ph.D. North Carolina (Chapel Hill) 1976. Aggregate wage dynamics and labor market disequilibrium: An errors-in-variables approach.

ROBERT D. WEAVER, Ph.D. Wisconsin (Madison) 1977. The theory and measurement of provisional production decisions.

LAURENCE M. WEISS, Ph.D. Harvard 1977. Risk, effort, and welfare theory.

ROBERT P. WITHINGTON, Ph.D. Pennsylvania 1976 Mutual savings banking in smaller urban centers of antebellum Massachusetts.

AKIRA YAMAZAKI, Ph.D. Rochester 1977. General equilibrium analysis of large nonconvex economies.

Economic Growth and Development; including Economic Planning Theory and Policy, Economic Fluctuations and Forecasting

BENJAMIN K. ACQUAH, Ph.D. Wisconsin (Madison) 1977. An analysis of demand for food commodities in the eastern region of Ghana.

MONA A. AL-BUSTANY, Ph.D. Catholic 1976. Trade and growth in the oil exporting economy of Iraq: Application of the two-gap model.

GEBEYEHU ALEMNEH, Ph.D. Michigan 1976. Development constraint in Ethiopia The next decade.

KOLAWOLE M. ALLI, Ph.D. Iowa State 1977. Production, income, and employment effects of agricultural technology in the central cocoa belt of western Nigeria: A multiperiod programming approach

ANNA L.O. DE ALMEIDA, Ph.D. Stanford 1977. Industrial subcontracting of low-skill service workers in Brazil.

DENISARD C. ALVES, Ph.D. Yale 1977. Manufacturing development in Ecuador: Dualism and x-efficiency.

EMMANUEL C. ANUSIONWU, Ph.D. Cornell 1977 A framework for a regional study of African agricultural rural economy· A case study of Kigezi district in western Uganda.

PRAKASH APTE, Ph.D. Columbia 1977. Intersectoral migration and economic growth.

MUHAMMAD M. AWAN, Ph.D. Clark 1976. Foreign capital and development process: The Pakistani experience.

MAHMOOD A. AYUB, Ph.D. Yale 1977. Income inequality in a growth-theoretic context: The case of Pakistan.

SURJIT S. BHALLA, Ph.D. Princeton 1977. Two essays· Savings and farm production in rural India.

TRIDIB K. BISWAS, Ph.D. Brown 1977. A theoretical study of the growth and development in an open dual economy.

CHARLES S. CALLISON, Ph.D. Cornell 1976. Land-to-tiller in the Mekong Delta: Economic, social, and political effects of land reform in four villages of South Vietnam.

DENIS G. CARTER, Ph.D. Florida 1976. Economic development and integration: A conceptual framework.

HENRI CAUVIN, Ph.D. New School 1977. The Haitian economy: A case study of underdevelopment.

ALFREDO CEBALLOS, D.B.A. Harvard 1977. Planning for economic development—a managerial approach A study of the process of export development.

THOMAS PEI-FAN CHEN, Ph.D. City (New York) 1976. Economic growth and structural change in Taiwan, 1952–72· A production function approach.

ERIC CHETWYND, JR., Ph.D. Duke 1976. City-size distribution, spatial integration, and economic development in developing countries. An analysis of some key relationships.

S. CHIA, Ph.D. McGill 1976. Industrialization strategy and industrial performance in Singapore, 1960–73

CHINYAMATA CHIPETA, Ph.D Washington (St. Louis) 1976. Family farm organization and commercialization of agriculture.

SUPOTE CHUNANUNTATHUM, Ph.D. Oregon 1977. An econometric analysis of demand and supply elasticities for Thai white rice exports

ENYINNA CHUTA, Ph.D. Michigan State 1977. Linear programming analysis of small scale industries in Sierra Leone.

CAROL A CORRADO, Ph.D. Pennsylvania 1976. The steady state and stability of the MIT-PENN-SSRC model and their implications for economic policy

ALAN S. COSTA, Ph D. California (Davis) 1977. Some developmental implications of alternative land tenure systems in the Mexican Pacific North.

ALFREDO J. DAMMERT, Ph.D. Texas (Austin) 1977. A world copper model for project design.

TRAN THANH DANG, Ph.D. Syracuse 1977. The distribution of income and wealth among individuals in the labor-surplus economy.

PAUL T. DAVENPORT, Ph.D. Toronto 1976. Capital accumulation and economic growth.

ADRIANO B. DIAS, Ph.D. Vanderbilt 1976. Market demand and income distribution.

ROBERT F. DIELI, Ph.D. Texas (Austin) 1977. Competition and efficiency in a developing capital market: The case of Mexico, 1966–73.

DAVID L. DORENFELD, Ph.D. Michigan 1977. Growth fluctuations in planned economies: A theoretical and econometric analysis.

SALEH H. EL MAIHUB, Ph.D. Colorado State 1977. Public investment in a capital-surplus country: The case of Libya.

BASHIR M. EL-WIFATI, Ph.D. Missouri (Columbia) 1977. Some socioeconomic considerations in the Bedouins' agricultural settlement: An example from Libya.

MOHAMMAD FAISAL, Ph.D. Michigan State 1977. Optimal land and water use and production response under alternative technologies in Bangladesh A programming approach

LISA P. FOX, Ph.D. North Carolina (Chapel Hill) 1976. Building construction as an engine of economic growth: An evaluation of the Columbian development plan.

EMANUEL A. FRENKEL, Ph.D. California (Davis) 1977. History and analysis of economic policy in De Gaulle's Fifth Republic

ALAN I. FRISHMAN, Ph.D. Northwestern 1977. The spatial growth and residential location pattern of Kano, Nigeria.

JULIO A. GENEL, Ph.D. Chicago 1977 On the state's strategy for financial development. The problem of noninflationary financing in Mexico

ROBERT GHOGOMU, Ph.D Northern Illinois 1976. Exports and economic development. The Cameroon experience

CLAUDIO GONZALEZ-VEGA, Ph.D Stanford 1976 On the iron law of interest rate restrictions Agricultural credit policies in Costa Rica and in other less developed countries

RICHARD L. GRABOWSKI, Ph D. Utah 1977 A socioeconomic theory of the development of agriculture

ARGHA GUHA, Ph D. McMaster 1976 The effects of alternative income distributions on resource allocation in India.

JAVAD HABIBION, Ph.D New School 1977. The impact of external resource inflows on the domestic savings of the developing countries. The case of Iran in the decade of 1962 72

HANS HANSEN, Ph.D. Pennsylvania 1977. The economy of Greenland, 1955-71

CHRISTOPHER J. HEADY, Ph D. Yale 1976 The determination of industrial wages in less developed countries.

ROBERT H. HENDRICKS, Ph.D Missouri (Columbia) 1976 Public land employed in different predominant uses: Its impact on the finances of the county court and school districts in Missouri

ALFRED J. HERSCHEDE, Ph.D. Illinois 1976. Investments in education and economic growth in the People's Republic of China.

WALTER HOPE, Ph.D. Catholic 1977. Opportunities and problems for development planning in a colonial setting, 1947-66: A case study of Guyana.

ANWARUL HOQUE, Ph.D. Michigan State 1977. A system simulation approach to policy planning and evaluation in the context of integrated rural development in Bangladesh.

EUI GAK HWANG, Ph.D. Oregon 1976. Demand for food and nonfood by farm and nonfarm households in Korea.

PARVIZ JENAB, Ph.D. Indiana 1977. The role of the third plan in economic development, Iran: A case study, 1963-67.

JAN JORGENSEN, Ph.D. McGill 1977. Structural dependence and economic nationalism in Uganda, 1888-1974.

ANA MARIA JUL, Ph.D. Pennsylvania 1977. A macroeconomic model for Brazil.

SALVADOR KALIFA-ASSAD, Ph.D. Cornell 1977. Income distribution in Mexico: A reconsideration of the distributive problem.

ROSTAM M. KAVOUSSI, Ph.D. Harvard 1976. Structural change in the Iranian manufacturing industry: 1959-72.

JACOB J. KETTOOLA, Ph.D Southern California 1977. Investment criteria in development planning, theory, practice, and policy

CHOEDCHAI KHANNABHA, Ph.D. Michigan 1977. Cost-benefit analysis of the Pa Mong project.

ROBIN D. KIBUKA, Ph.D. Harvard 1977. Income distribution and fiscal incidence in Uganda, 1961-70.

DAEMO KIM, Ph D. Rice 1976. Structural change, employment, and income distribution: The case of Korea, 1960-70.

RONALD KIZIOR, Ph.D. Notre Dame 1976. The effects of government contracts on regional growth and employment: A case study of the SMEK area

JOSEPH D. LAFORTE, Ph D. Connecticut 1977 Water pollution control costs and public policy for the paper industry with application to western Massachusetts.

EZZEDDINE LARBI, Ph.D California (Los Angeles) 1976. Foreign capital inflow and optimal external indebtedness for the Tunisian economy: The application of control theory to policy problems

CHUL-HEUI LEE, Ph.D. Harvard 1977. Export promotion and economic development in Korea: 1962-73

JEFFREY E. LEVIN, Ph.D. Purdue 1976. Heterogeneous labor, disguised unemployment, and economic growth

ROBERT LEY, Ph.D. Washington State 1977. Bolivian tin and Bolivian development.

DEAN A. LINSENMEYER, Ph.D Michigan State 1976. Economic analysis of alternative strategies for the development of Sierra Leone marine fisheries.

ADEWALE MABAWONKU, Ph.D. Michigan State 1977 The role of apprenticeship training in the small-scale industrial subsector of western Nigeria.

BEKHA L. MAHARAJAN, Ph.D. Missouri (Columbia) 1976. An evaluation of agricultural production systems and alternatives for small farms in Nepal.

ANTONIO C. MARTIN DEL CAMPO, Ph.D. California (Berkeley) 1976. Programming Mexican agricultural change recognizing the heterogeneous agrarian structure. A case study of Nayarit.

MOTHAE MARUPING, Ph.D. Catholic 1977. The re-examination of the international demonstration effect on the economies of less developed countries: With special reference to southern Africa.

WILLIAM B MARXSEN, Ph.D. Georgia State 1976. A monetary and fiscal policy in a growing economy.

PETER J MATLON, Ph.D. Cornell 1976 The size distribution, structure, and determinants of personal income among farmers in the north of Nigeria.

EILEEN MAUSKOPF, Ph D. Johns Hopkins 1977. Inflation expectations in Israel. Direct estimates from purchasing-power bonds.

GINIGEMF F. MBANEFOH, Ph D. Illinois 1976 Allocation of road funds in Nigeria: An evaluation.

PEDRO C. DE MELLO, Ph.D. Chicago 1977 The economics of labor in Brazilian coffee plantations.

JOZE MERCINGER, Ph D. Pennsylvania 1975. A quarterly econometric model of the Yugoslav economy

GARY F MILLER, Ph.D. Southern California 1977 Selected aspects of full employment policies in France, West Germany, and the United Kingdom, 1955 69

JEFFREY B MILLER, Ph D Pennsylvania 1976 National planning in a market system Some normative considerations.

RAKESH MOHAN, Ph D Princeton 1977. Development, structural change, and urbanization: Explorations with a dynamic three-sector general equilibrium model applied to India, 1951-84.

AHMAD MOJTAHED, Ph D Iowa State 1977 Economics of livestock development in Iran

DOW MONGKOLSMAI, Ph.D Cornell 1977 Distributional effects and reimbursement analysis of an irrigation project in Thailand

TAHANY R NAGGAR, Ph D Oklahoma 1976. The role of the iron, copper, lead, and zinc petroleum export sectors in the economic development of Brazil, Chile, Mexico and Venezuela. An empirical macro-economic analysis.

SALIH N. NEFTCI, Ph D Minnesota 1977. Three essays in business cycle research.

EBERHARDT V. NIEMEYER III, Ph.D. Texas (Austin) 1976. The effect of energy supply on economic growth

MICHAEL P. O'NEILL, Ph D Oklahoma 1977. An analysis of entrepreneurship in economic development: A synthesis of Schumpeter, Hagen, and McClelland.

OBIDIMA I.O. ONYEMELUKWE, Ph.D Michigan 1977. Econometric analysis of structural patterns of economic development.

GUILLERMO ORTIZ-MARTINEZ, Ph D Stanford 1977. Capital accumulation and economic growth: A financial perspective on Mexico.

WILLIAM J. OYAIDE, Ph.D. Temple 1976. The role of direct private foreign investments in economic development: A case study of Nigeria, 1963-73.

CHI-DO PHAM, Ph.D. Pennsylvania 1976 Inflationary finance in wartime South Vietnam, 1960-72.

DIBYO PRABOWO, Ph.D. Washington State 1977. Water development and farm production possibilities in the Solo River Basin of Indonesia: A linear programming analysis.

PHILIP K. QUARCOO, Ph.D. Western Ontario 1976. Efficient credit allocation in Ghana.

MAHMOOD A. RAZ, Ph.D. Stanford 1977. Structural changes in the labor force and economic development in India

PEDRO A REYES-ORTEGA, Ph D. Southern California 1977. A macroeconomic model for Mexico with emphasis on income distribution.

STUART K SHWEDEL, Ph D Michigan State 1977 Marketing problems of small farm agriculture: Case study of the Costa Rican potato market

CONSTANTINOS C. SIFNIOTIS, Ph.D. Northeastern 1976 Education and the patterns of consumption: Their impact on the demand for highly educated labor.

ALVARO SILVA, Ph D. Michigan State 1976. Evaluation of food market reform Corabastos-Bogota.

IRLAN SOFJONO, Ph D. Iowa State 1977 Growth and distributional changes of paddy farm income in central Java, 1968-74.

KUTLU SOMEL, Ph.D. Stanford 1977. Economics of improved dryland wheat technology. A case study of Ankara, Turkey

JAMES E. STEINER, Ph D. Georgetown 1977 Economic methods for military threat assessment.

JOHN SWEENEY, Ph.D. Catholic 1977. An economic analysis of the nationalization of the Gran Mineria of copper in Chile.

TAT-WAI TAN, Ph.D. Harvard 1977 Income distribution and determination in West Malaysia.

ABUBAKAR USMAN, Ph.D. Minnesota 1976. The Phillips tradeoff relation in Nigeria: Presumptive exploitability, institutional reforms/policy operations, a theoretical explanation and some evidence.

MOHAMMAD R. VAEZ-ZADEH, Ph.D. Johns Hopkins 1977. Optimal economic growth with exhaustible resources and absorptive capacity constraint: The case of Iran.

RENE P. VILLARREAL-ARRAMBIDE, Ph.D. Yale 1976 External disequilibrium and growth without development: The import substitution model (The Mexican experience, 1929-75).

RAMESH R. WAGHMARE, Ph.D. McMaster 1976. The portfolio behavior of industrial corporations in India.

MICHAEL T. WEBER, Ph.D. Michigan State 1976. An analysis of rural food distribution in Costa Rica.

ALLAN N. WILLIAMS, Ph.D. Cornell 1976. Agricultural reorganization and the economic development of the working class in Jamaica.

DAVID WILLIAMS, D.B.A. Harvard 1976. Choice of technology and national planning: The case of Tanzania.

GEORGE E. WRIGHT, JR., Ph.D. Michigan 1977. Regional inequality in the economic development of Iran, 1962–70.

JESUS YANEZ-ORVIZ, Ph.D. Michigan State 1976. Optimal allocation of housing investment in five Mexican cities, 1960–70 and 1970–85.

HA-CHEONG YEON, Ph.D. City (New York) 1977. Estimation of the Cobb-Douglas and *CES* production functions in Korea, 1957–75.

HASSAN A. ZAVAREEI, Ph.D. New School 1977. Dependent industrialization in Brazil: Including a case study of the motor vehicles industry.

Economic Statistics; including Econometric Methods, Economic and Social Accounting

PETER C. BELL, Ph.D. Chicago 1977. A capacity expansion model with economies of scale.

NILS S. BLOMQUIST, Ph D. Princeton 1976 The distribution of lifetime income A case study of Sweden

STEVEN D. BRAITHWAIT, Ph.D. California (Santa Barbara) 1976. Consumer demand and cost of living indexes for the United States An empirical comparison of alternative multilevel demand systems.

SHIRLEY CASSING, Ph.D. Iowa 1976. Contributions to autocorrelation theory.

STEPHEN J CHAZEN, Ph D California (Los Angeles) 1977. On the predictability of the p-class estimator in economic models.

KE-YOUNG CHU, Ph D Columbia 1976. The dynamic simultaneous equations model with moving average errors.

FRANK COOPER, Ph.D Wisconsin (Madison) 1977. On the dynamic properties of the MPS model: A simulation study.

PHILIP G. ENNS, Ph.D. Michigan 1976. A Bayesian approach to Kalman varying parameter time-series estimation with business applications.

DENNIS E. FARLEY, Ph.D. Pennsylvania 1976. A study of the Hannan inefficient estimator.

JAMES R. FROESCHLE, Ph.D. Iowa 1975. Analysis of seasonal time-series.

GAMINI D. GUNAWARDANE, Ph.D. Chicago 1977 Implicit representation of special structure constraints in the linear programming solution of some production-distribution problems.

SUNG SHIN HAN, Ph.D. Pennsylvania 1977. Three essays on application of optimal control to econometric models.

HIDEO HASHIMOTO, Ph.D. Illinois 1977. World food projection models, projections, and policy evaluation.

WING-CHUNG HO, Ph.D. Vanderbilt 1976. Specification analysis and non-linear estimation of *CES* production functions.

ABRAM E. HOFFMAN, D.B.A. Harvard 1976. Limited bandwidth distributed lags.

HAE-SHIN HWANG, Ph.D. Minnesota 1976. An econometric analysis of single-family, owner-occupied housing market

RAPHAEL JUANTORENA, Ph.D. Louisiana State 1977. Application of a modified Kalman filter algorithm to economic estimation and identification.

TERRACE W KINAL, Ph.D. Minnesota 1976. The exact finite sampling distribution of two-stage least squares estimates

BETSEY A. E. KUHN, Ph.D. Stanford 1977. An estimation model for futures contract margin requirements.

JUNG SOO LEE, Ph.D. State University of New York (Albany) 1977. Econometrics with unobservable variables: Four essays in monetary economics.

LUNG-FEI LEE, Ph.D Rochester 1977. Estimation of limited dependent variable models by two-stage method.

WINSTON LIN, Ph.D Northwestern 1976. Econometric factor analysis

LAURENCE A. MADEO, Ph.D. Michigan 1976 A formalization of multidimensional search for organizational planning through terminal based models.

AN-SIK MIN, Ph.D. Iowa 1975. A study of likelihood functions of ARIMA autoregressive integrated moving average processes

CARL J PALASH, Ph D. Pennsylvania 1976. The optimal control of the MPS econometric model.

PEDRO A PALMA-CARILLO, Ph.D. Pennsylvania 1976. A macroeconometric model of Venezuela with oil price impact applications.

CHARLES I. PLOSSER, Ph.D. Chicago 1976. A time-series analysis of seasonality in econometric models with an application to a monetary model.

CHARLES G. RENFRO, Ph.D Pennsylvania 1976. A quarterly econometric model of the state of Kentucky.

MATTHEW J. ROBERTSON, Ph.D. Queen's 1977. The small sample properties of several limited information estimators in simultaneous equation models with first-order autoregressive errors and lagged dependent variables.

DEXTER R. ROWELL, Ph.D. Pennsylvania 1975. Macro system estimation and its implication for prediction efficiency.

RAVI SARATHY, Ph.D. Michigan 1976. Location-allocation problems in multinational firms: A computer simulation model.

HAROLD J. SCHLEEF, Ph.D. Chicago 1977. Optimal control models for multiserver exponential queueing systems.

ROBERT SCOTT, Ph.D. Iowa 1975. Smear and sweep: A method of forming indices for use in testing in non-linear systems.

ROBIN C. SICKLES, Ph.D. North Carolina (Chapel Hill) 1976. Simultaneous equations models containing truncated endogenous variables.

GARY R. SKOOG, Ph.D. Minnesota 1976. Two essays in time-series analysis.

ROBERT K. SPRINGER, Ph.D. Illinois 1976. Pooling time-series and cross-section data in a system of simultaneous equations.

THOMAS C. WHITEMAN, Ph.D. Pennsylvania State 1976. The effects of the business cycle on the size distribution of labor incomes.

JIEH-JOU WU, Ph.D. Southern Methodist 1977. The demand for automobile service with an emphasis on waiting interval: A cross-section study.

ANNIE ON LAI YUEN, Ph.D. Wisconsin (Madison) 1977. Dynamic modeling in an agricultural subsector: An analytic-empirical approach.

Monetary and Fiscal Theory, Policy, and Institutions

ALI A. ABDUSSALAM, Ph.D. Cincinnati 1976. Money supply in a small open economy. The case of Libya.

ELIZABETH A. BAHLKE, Ph.D. Michigan 1976. Individual bank demands for certificate of deposit funds

CARLISS Y. BALDWIN, D.B.A Harvard 1977. Illiquid assets and liquidity preference in an uncertain environment.

LESTER BARENBAUM, Ph.D. Rutgers 1977. Bank performance and market structure: The case of New Jersey.

M. ALAN BAUGHCUM, Ph.D. North Carolina (Chapel Hill) 1976. The federal highway program: A case study in fiscal federalism.

JOHN BECK, Ph.D. Michigan State 1976. State aid and equal education opportunity in Michigan.

NIELS BLOMGREN-HANSEN, Ph.D. Pennsylvania 1975. An econometric study of the financial sector in Denmark.

JAMES L. BUTKIEWICZ, Ph.D. Virginia 1977. Some macroeconomic effects of government debt.

RALPH T. BYRNS II, Ph.D. Rice 1977. Compositional instability and the attainment of macroeconomic goals.

JAMES T. CAMPEN, Ph.D. Harvard 1976. Public expenditure analysis: A critical analysis of current theory and practice together with some contributions toward a participatory alternative.

WILLIAM H. CARLSON, Ph.D. Carnegie-Mellon 1977. An investigation of fiscal and monetary policy on the economy: 1922-40; 1947-70; 1971-75.

KEITH CARPENTER, Ph.D. Wisconsin (Madison) 1976. Irving Fisher's monetary theory: An evaluation.

DOMINGO F. CAVALLO, Ph.D. Harvard 1977. The stagflationary effect of monetary stabilization policy in economies with persistent inflation.

RONALD C. CLUTE, Ph.D. Notre Dame 1977. An analysis of the incidence of federal sharing funds in local government.

GEORGE R. COMPTON, Ph.D. California (Los Angeles) 1976. The economics of discretionary collective goods.

JEAN M. DEBOIS, Ph.D. California (Berkeley) 1976. Policy choice, voting, and prisoners' dilemma game, with an application to EEC agriculture support.

DONAL J. DONOVAN, Ph.D. British Columbia 1977. Consumption, leisure, and the demand for money and money substitutes.

SOL DRESCHER, Ph.D. City (New York) 1977. The demand for money: A household production function approach.

MARTIN M. EBNER, Ph.D. Florida 1976. Development, estimation, and forecasting accuracy of regional financial models: An application within the state of Florida.

ROBERT C. EVANS, Ph.D. Washington (St. Louis) 1977. Information technology and stock market organization.

JAMES A. FELLOWS, Ph.D. Louisiana State 1977. Some welfare implications of the regulation of commercial bank entry and branching.

RONALD C. FISHER, Ph.D. Brown 1977. A theoretical analysis of general revenue sharing effects on subordinate government public expenditures.

JAMES R. FOLLAIN, Ph.D. California (Davis) 1976 Estimating the impact of grants upon local fiscal behavior: Does econometric methodology and collective choice specification really make a difference?

CHARLES T. FRANCKLE, Ph.D. North Carolina (Chapel Hill) 1977. An analysis of household sector demand for financial assets, 1952-75.

PETER GARBER, Ph.D. Chicago 1977. Costly decisions and the demand for money.

BETTY GIBSON, Ph.D. Iowa 1976. Measurements of preference variation for indirectly consumed goods and the association of this measured variation with tax and expenditure preferences.

RAE J. B. GOODMAN, Ph.D. Washington (St. Louis) 1976. Uncertainty and liability structure in portfolio analysis.

JO ANNA GRAY, Ph.D. Chicago 1976. Essays on wage indexation.

JACOB GROSSMAN, Ph.D. Columbia 1976. Anticipated and unanticipated nominal income growth, and short-run fluctuations in unemployment and policies in the United States.

WILLIAM C. GRUBEN, Ph.D. Texas (Austin) 1977. A regional approach to some applications of multi-

variate statistical analysis to the assessment of bank performance.

HARLAN I. HALSEY, Ph.D. Stanford 1976. Implicit income taxation in the negative tax system.

WILLIAM R. HART, Ph.D. Washington (St. Louis) 1976. The financing constraint and macroeconomic analysis.

BRYON M. HIGGINS, Ph.D. Michigan 1977. The microeconomic determinants of the demand for demand deposits.

WILLIAM M. HILDRED, Ph.D. Colorado State 1976. Passive tax expenditures.

MICHAEL W. JOHNSON, Ph.D. Northwestern 1977. A pure theory of local public goods.

MARCOS T. JONES, Ph.D. Princeton 1977 Three essays on portfolio adjustment.

GEORGE M. KATSIMBRIS, Ph.D. Connecticut 1977. Size distribution effects and the business demand for money: An econometric study.

DAVID C. KAVANAUGH, Ph.D. Claremont 1977. Interest rate risk and expectations An application to the maturity structure of interest rates

IRA G. KAWALLER, Ph D. Purdue 1976. The dynamics of the housing market, with special emphasis on the impact of financial conditions and housing subsidies.

MICHAEL E. KENNEDY, Ph.D. Queen's 1976. Monetary policy in a monetarist framework The Canadian experience from 1963 to 1969.

ARTHUR T. KING, Ph.D. Colorado 1976 The relationships between socioeconomic programs and the Department of the Air Force budget: Section 8(a) of the Small Business Act – the economic development and public finance aspects of the public policy program.

TIMOTHY W. KOCH, Ph.D. Purdue 1976. An econometric model of the market for tax-exempt securities.

GEORGE P. LEPHARDT, Ph.D Tennessee 1976. A six-country comparison of money supply changes, and fiscal expenditures on *GNP*.

ANDREA LUBOV, Ph.D. Washington State 1977. Financing Washington's public schools: Equity considerations.

ARTHUR LYONS, Ph.D Northwestern 1977. An economic theory of campaign contributions

PATRICIA A. MCGUIRE, Ph.D. California (Los Angeles) 1977. Some efficiency characteristics of local government.

PRANLAL MANGA, Ph.D. Toronto 1976. A benefit incidence analysis of the public medical and hospital insurance programs in Ontario

PAUL L. MENCHIK, Ph.D. Pennsylvania 1976. A study of inheritance and death taxation: A microeconometric approach.

NELSON C. MODESTE, Ph.D. Florida 1976. Economic instability and the monetary process of adjustment in an open economy.

STEPHEN O. MORRELL, Ph.D. Virginia Polytechnic Institute 1977. Monetary arrangements and media of exchange: An analysis of the effects of the New Deal monetary reform.

JAY B. MORRISON, Ph.D. California (Berkeley) 1977. Interest rate risk in commercial banking: Some implications for capital adequacy.

BRENDA M. NADUCKA, Ph.D. Notre Dame 1977. An examination of tax and expenditure incidence by neighborhood: A study of South Bend, Indiana.

HAROLD N. NATHAN, Ph.D. California (Davis) 1977. Monetary aspects of the bond pegging period 1942-51.

SANDRA J. ODORZYNSKI, Ph.D. Purdue 1976. Allocation, distribution, and welfare costs from tax-induced distortions of labor force participation rates.

JOSE T OLIVEIRA, Ph D Purdue 1976. Tax on industrialized products: A case study on value-added taxation.

BENEDICT J. PENDROTTI, Ph.D. Michigan 1977. An aggregate model of member bank portfolio adjustment.

RICCARDO R. PELLICCIARO, Ph.D. Fordham 1977. A test of the efficient capital market model.

GEORGE M. PERKINS, Ph.D Boston College 1977. Demand for public goods.

PHILIP W PERRY, Ph.D. Stanford 1977. Money demand, bank profits, and expected inflation.

DENNIS PESEAU, Ph.D. Claremont 1977. A specification of a demand for liquid assets.

WAINO PIHL, Ph.D. Wayne State 1977. The money demand function: A test for the proper specification and the existence of short-run stability.

GLENN T. POTTS, Ph.D. Iowa State 1976 An inquiry into the macroeconomic objectives of monetary policy, 1956-75.

PAUL H. RACHAL, D.B.A Harvard 1977. Commercial bank loan pricing policy.

MICHAEL J RILEY, D.B.A Harvard 1977 A study of NOW account strategies The Massachusetts and New Hampshire experience

JOANNA F ROBINSON, Ph.D Connecticut 1977 A new look at costs and benefits of membership in the Federal Reserve System.

CHARLOTTE E. RUEBLING, Ph.D. Iowa State 1976. An explanation of the behavior of the ratio of time deposits to demand deposits.

MICHAEL K. SALEMI, Ph D. Minnesota 1976. Hyperinflation, exchange depreciation, and the demand for money in post-World War I Germany.

JOAO SAYAD, Ph.D. Yale 1976. Regulation of Brazilian commercial banks.

L. WAYNE SHELL, Ph D Louisiana State 1977. Reve-

nue growth and a state's tax structure: The case of Louisiana.

ALDEN F. SHIERS, Ph.D. California (Santa Barbara) 1977. Mortgage commitments and the supply of single-family housing starts.

KI-RYON SHIM, Ph.D. Cincinnati 1976. The shifting of the corporate income tax in Japanese manufacturing.

A. FREDERICK SIEGMUND, Ph.D. Oklahoma State 1976. A spatial approach to government.

JOSEPH A. SONNEMAN, Ph.D. Claremont 1977. Decision making in public finance.

TEIZO TAYA, Ph.D. California (Los Angeles) 1977. A study on the movement of stock prices in the Tokyo stock exchange.

ERNANI TEIXEIRA, Ph.D. California (Los Angeles) 1976. Money demand, money supply, and prices in Brazil, 1950–72.

ASHER TISHLER, Ph.D. Pennsylvania 1976. Econometric model of the commercial banks sector in a complete flow of funds model of the United States.

LEONARD L. TSUMBA, Ph.D. Virginia Polytechnic Institute 1977. An analysis of monetary velocity in an open economy.

EURICO H. UEDA, Ph.D. Vanderbilt 1976. An analysis of the effects of the 1967–71 Brazilian fiscal reform.

CHARLES L. VEHORN, Ph.D. Ohio State 1977. Estimating cross elasticities between public and private goods.

JOHN R. WAGNER, Ph.D. Temple 1976. The relative efficacy of monetary and fiscal policy in macroeconomic models: An analysis of the determinants of the relative potency, speed, and stability of the multipliers.

DONALD A. WATNE, Ph.D. California (Berkeley) 1977 Inferences about the structural redesign of a disaggregated branch banking system having decomposition externalties.

DAN E. WATSON, Ph.D. Rice, 1977 Interest rate expectations in commercial bank portfolio behavior.

BETSEY R. B. WHITE, Ph.D Stanford 1977. On the rationality of observed saving. A critique of the life cycle hypothesis.

DAVID WILDASIN, Ph.D. Iowa 1976. Theoretical issues in local public finance.

T. HINDS WILSON, Ph.D. North Carolina (Chapel Hill) 1976. The permanent rate of investment and lags in monetary policy.

RONALD S. WINTROBE, Ph.D. Toronto 1977. The economics of bureaucracy.

VICHIT WONGVASU, Ph.D. Utah 1977. The determinants of the money supply in Thailand.

CYNTHIA WOOD, Ph.D. Wisconsin (Madison) 1976. Nonbank financial intermediation and the effectiveness of monetary policy.

JAMES N. YOUNG, Ph.D. Georgia State 1976. The implementation requirements and stabilization properties of monetary policy guided by feedback control rules.

ASGHAR ZARDKOOHI, Ph.D. Virginia Polytechnic Institute 1977. On the performance of public enterprises: The choice of price and productive factors.

LEONARD V. ZUMPANO, Ph.D. Pennsylvania State 1976. The quantitative dimensions of municipal income taxation.

International Economics

ALI I. ABDI, Ph.D. Washington (St. Louis) 1976 Banking, capital formation, and economic development: A comparative approach with emphasis on the eastern African experiences.

SAEED ABTAHI, D.B.A. Harvard 1976. Financial flows to developing countries: The case of Eurodollar credits.

STUART D. ALLEN, Ph D. Virginia 1977. The causation of inflation in Switzerland: 1952–75.

ANNAN AMEGBE, Ph.D Catholic 1977. An analysis of macroeconomic effectiveness of foreign capital in a developing country A case study of Ghana, 1957–65.

JAMES L. BEAVER, Ph.D. Virginia 1977. Monetary policy and the money supply of Sweden

RICHARD B. BERNER, Ph D. Pennsylvania 1976. An empirical general equilibrium model of international discrimination

JOHN F.O. BILSON, Ph.D. Chicago 1976. A monetary approach to the exchange rate.

BASUDEB BISWAS, Ph.D Chicago 1976. An economic analysis of India's export performance, 1950–70.

GERARD CAPRIO, JR., Ph.D. Michigan 1976 The international transmission of inflation in the short run.

JAMES H. CASSING, Ph.D. Iowa 1975. Transport costs and the pure theory of international trade.

KENNETH S. CHAN, Ph.D Brown 1977. Exchange rates and tariffs in a nonmarket clearing model of international finance.

DORIS GUANG-MEEI CHENG, Ph.D. Notre Dame 1976 An empirical study of the effects of U S grain exports on the exchange rate and domestic prices.

KENNETH W. CLEMENTS, Ph.D. Chicago 1977. The trade balance in monetary general equilibrium.

JAMES H. COBBE, Ph.D. Yale 1977. Some aspects of government policy towards foreign extractive enteprises in less developed countries.

LUIS L. CONDE, Ph.D. Boston College 1977. Imported vs. policy-induced prices and payments in the Dominican equilibrium.

WILLIAM M. COX, Ph.D. Tulane 1976. Rational expectations in an open economy.

MICHAEL J. CROSSWELL, Ph.D. Northwestern 1976. Optimal policy in an open economy with variable endogenously determined distortions.

BRUCE R. DALGAARD, Ph.D. Illinois 1976. South Africa's impact on Britain's return to gold, 1925.

BETTY C. DANIEL, Ph.D. North Carolina (Chapel Hill) 1976. Analysis of stabilization policies in open economics under conditions of inflation and inflationary expectations.

ALEYA DARMOUL, Ph.D. New York 1977. Customs union, monetary union, and economic integration in North-West Africa.

SANDWIP K. DAS, Ph.D. Southern Methodist 1976. The theory of international trade under terms of trade uncertainty.

ROBERT R. DAVIS, Ph.D. Virginia Polytechnic Institute 1977. Eurocurrency markets and control of monetary aggregates.

LIVIO W.R. DECARVALHO, Ph.D. Cornell 1977. Comparative performances of domestic and foreign firms in Latin America.

GLENN R. DESOUZA, Ph.D. Fordham 1976. A simultaneous equation model of the *U.S.* merchandise trade account.

MARK R. EAKER, Ph.D. Stanford 1977. Hedging and the theory of forward exchange.

JOHN C. EDMUNDS, D.B.A. Harvard 1977. Strategies of Latin American exporters.

IKUSHI EGAWA, Ph.D. Rochester 1977. Existence and comparative statistics of world trade equilibrium.

DONALD G. FERGUSON, Ph.D. Toronto 1976. Some essays on international factor mobility and the theory of comparative advantage.

VINCENT W.J. FITZGERALD, Ph.D. Harvard 1976. The inflationary process in an open economy: Canada 1961-74.

ROBERT P. FLOOD, Ph.D. Rochester 1977. Essays on a monetary approach to real and financial aspects of various exchange rate systems.

KENNETH C. FROEWISS, Ph.D. Harvard 1977. An analysis of international yield curve differentials.

JORGE GARCIA-GARCIA, Ph.D. Chicago 1976. Foreign trade and monetary-fiscal policies in Colombia: 1953-70.

WILLIAM J. GASSER, Ph.D. Ohio State 1976. The Canadian money supply under fixed and flexible exchange rates.

LANCE GIRTON, Ph.D. Chicago 1976. Monetary policy in the open economy.

MARC GOLD, Ph.D. Wayne State 1976. The 1965 Canadian-American automotive trade agreement: An econometric evaluation.

ANTONIO GOMEZ-OLIVER, Ph.D. Chicago 1977. The adjustment of the money market, the price level, and the balance of payments in Mexico.

ROBERT E. GROSSE, Ph.D. North Carolina (Chapel Hill) 1977. Foreign investment codes and the location of international investment.

DONALD M. HARRISON, Ph.D. Virginia 1977. Monetary policy in an open country: The case of Belgium, 1951-74.

DAVID G. HARTMAN, Ph.D. Harvard 1976. Taxation of foreign-source investment income.

STEPHEN E. HAYNES, Ph.D. California (Santa Barbara) 1976. International capital markets: A frequency domain analysis within portfolio and Fisherian frameworks.

CHRISTINE R. HECKMAN, Ph.D. Chicago 1977. Structural change and purchasing power parity.

DIRK HEREMANS, Ph.D. California (Los Angeles) 1977. Monetary policy and international capital flows in small open economies: The Belgian experience, 1961-73.

ROBERT HODRICK, Ph.D. Chicago 1976. The monetary approach to the determination of the exchange rate: Theory and empirical evidence.

CHUNG-YUAN HSU, Ph.D. Rice 1976. Export promotion in Taiwan: An assessment.

KENT H. HUGHES, Ph.D. Washington (St. Louis) 1976. International economic decision making in Congress: A case study of the Burke-Hartke bill.

JOONG-KI HWANG, Ph.D. Utah 1977. Determinants of foreign investment with special reference to Japanese investment in Korean industry.

MICHAILO IVANOVIC, Ph.D. Columbia 1977. A model for policy analysis within the European Economic Community.

JANICE W. JADLOW, Ph.D. Oklahoma State 1977. Trade liberalization in the chemical industry: The impact on the United States.

JOHN E. JELACIC, Ph.D. North Carolina (Chapel Hill) 1976. Pollution control and the international trade of aluminum: A multilateral study of the changes in the trade flows of the aluminum industry.

SCOTT T. JONES, Ph.D. Virginia Polytechnic Institute 1977. A variable risk hypothesis and the forward exchange market.

KU-HYUN JUNG, Ph.D. Michigan 1976. Effects of investment incentives on the profitability of foreign investment projects.

LAWRENCE L. KREICHER, Ph.D. Harvard 1977. International capital flows and the balance of payments: Empirical studies of six European countries and the United States.

LEROY O. LANEY, Ph.D. Colorado 1976. Applications of portfolio choice theory to the post-Bretton Woods international monetary environment.

DALE W. LARSON, Ph.D. Wisconsin (Madison) 1976. Increasing returns to scale and international trade in similar manufactured goods.

JULES E. LE BON, Ph.D. Tulane 1976. The real and nominal effects of devaluation: The relative price effects of exchange rate movements.

CHI-WEN J. LEE, Ph.D. Rochester 1977. Money, prices, and balance of payments of Taiwan, 1952–74: A study of a small open developing economy under fixed exchange rate.

CHON PYO LEE, Ph.D. Brown 1977. A study on the optimal exchange market operation.

RICHARD M. LEVICH, Ph.D. Chicago 1977. The international money market: Tests of forecasting models and market efficiency.

WAYNE LEWIS, Ph.D. Wisconsin (Madison) 1977. The effects of multilateral trade liberalization on *U.S.* domestic prices.

YEANG-ENG LIM, Ph.D. Wisconsin (Madison) 1977 Import instability: An empirical analysis.

SUSAN E. MCGOWAN, Ph.D. California (Davis) 1977. A static and dynamic linear expenditure model applied to *U.S.* import demand.

SYED I. MAHDI, Ph.D. Massachusetts 1976 The effect of a free trade policy on the factor-intensity of United States-Japanese trade.

NANCY P. MARION, Ph.D. Princeton 1977 Two-tier exchange rates and the role of monetary policy.

THOMAS MATHEW, Ph D. Georgia 1976. Economics of international cartels: A case study of sugar.

SANGKEE MIN, Ph.D Michigan 1976. Determinants of interest differentials between the *U S* and Eurodollar markets: An empirical investigation.

JUAN H. MOLDAU, Ph.D. Vanderbilt 1976. Cost benefit analysis from international viewpoint

DOUGLAS W. MORRILL, Ph.D Indiana 1976. Economic effects of Japan's barriers to international trade and investment, 1951–73

CYRIL L MUNDELL, Ph.D North Carolina (Chapel Hill) 1976. Devaluation, capital movements, and short-run income redistribution· A two-country model.

RICHARD E. NELLIS, Ph.D. Pennsylvania State 1977. Determinants of direct private foreign investment in the manufacturing sector of the *U.S.* food system.

KAMRAN NOMAN, Ph D. Rochester 1977 Output, employment, and the balance of payments: Related essays on the process of adjustment in a small open economy.

JOSE A. OCAMPO-GAVIRIA, Ph D. Yale 1976. Capital accumulation and international relations.

PATRICIA L. PHELPS, Ph.D. Oklahoma 1977. The effect of debt-servicing and aid-tying on the grant element in foreign economic assistance.

NED H. PHILLIPS, Ph.D. Ohio State 1976. The welfare effects in the United States of removing import restrictions on beef.

JOHN G. POMERY, Ph.D. Rochester 1977. International trade and uncertainty: Simple general equilibrium models involving randomness.

BLUFORD H. PUTNAM, Ph.D. Tulane 1976. International price and interest rate differentials and the monetary approach to the balance of payments.

DAVID H. RESLER, Ph.D. Ohio State 1977. A general equilibrium model of the Eurodollar market.

LEE A. REYNIS, Ph.D. Michigan 1976. The proliferation of *U S* firm Third World offshore sourcing in the mid-to-late 1960's: An historical and empirical study of factors which occasioned the location of production for the *U.S* market abroad.

RAYMOND RIEZMAN, Ph.D. Minnesota 1977. A theoretical model of customs unions.

MICHAEL R. ROSENBERG, Ph.D. Pennsylvania State 1976. German monetary policy under alternative exchange rate systems

HIROHISA SAITO, Ph.D. Georgia State 1976. A study of the financial source of the outflow of *U.S.* manufacturing direct investment.

RICARDO L SANTIAGO, Ph D. Vanderbilt 1976. An application of international trade theories to interregional trade in Brazil.

CYRUS SASSANPOUR, Ph.D Brown 1977. The role of money in open economies· A theoretical and empirical analysis of the monetary approach to the balance-of-payments theory

AVINASH C. SETH, Ph.D. Southern Methodist 1976 Three essays on returns to scale and the two-sector model

ALAIN H SHEFR, Ph.D Duke 1976 The welfare costs of tariffs, transfers, and exchange rate destabilization.

HYUNCHUL SHIN, Ph.D. Utah 1977. Monetary policy in an open economy with particular reference to the recent Korean experience.

SOEDIJONO, Ph.D. Colorado 1977. Indonesian exchange control systems, 1950–65 A theoretical appraisal.

KOMAI S SRI-KUMAR, Ph.D. Columbia 1976 Tests of international interdependence in the monetary approach to the balance of payments.

JOE A STONE, Ph.D. Michigan State 1977. Price elasticities and the effects of trade liberalization for the United States, the European Economic Community, and Japan.

JOSEPH TEMPLEMAN, Ph D. City (New York) 1977. An empirical analysis of Israel's demand policy response to her balance-of-payments situation, 1956–73

JAN TER WENGEL, Ph.D. North Carolina (Chapel Hill) 1977. Allocation of industry in the Andean common market.

RICHARD THORNTON, Ph.D. Northern Illinois 1976. An econometric model of *U S.* short-term international capital flows: A multiple portfolio approach.

TSENG-HU TSAO, Ph.D. Pennsylvania 1976. Theoretic,

econometric, and input-output analyses of Taiwan's foreign trade.

PHILIP P. TURNER, Ph.D. Harvard 1976. Some effects of devaluation: A study based on the *U.K*'s trade in manufactured goods.

CARL VAN DUYNE III, Ph.D. Stanford 1977. Export controls and commodity prices.

MYLES S. WALLACE, Ph.D. Colorado 1976. The monetary approach: An examination of some theoretical and empirical aspects of flexible exchange rates: Canada-United States 1950 IV-1961 I.

YEN-KYUN WANG, Ph.D. Pennsylvania 1975. A macroeconomic analysis of the balance of payments of Korea and policy implications.

NANCY N. WARDELL, D.B.A. Harvard 1977. *U.S.* iron ore imports: Sourcing strategies for *U.S.* steel companies.

BERNARD M. WASOW, Ph.D. Stanford 1977. Policy in a small open economy with wage dualism.

ELIZABETH WEBBINK, Ph.D. New York 1977. *U.S.* foreign trade in manufactured goods in 1966-70, and the structure of the domestic market

CLAS G. WIHLBORG, Ph.D Princeton 1977. Risks, capital market integration, and monetary policy under different exchange rate regimes.

DWAINE S. WILFORD, Ph.D Tulane 1976 The open economy in a monetized world: The Mexican case.

LARRY N WILLMORE, Ph.D Carleton 1977. The impact of integration on the industrialization of Central America.

JOHN R. WOODBURY III, Ph.D. Washington (St Louis) 1977 Foreign investment A synthesis of industrial organization, trade, and capital theoretic approaches.

PHOT YONGSKULROTE, Ph.D Oklahoma State 1976. Exchange rates as a determinant of the balance of payments

Business Administration; including Business Finance and Investment, Insurance, Marketing, and Accounting

ROBERT M. ALLOWAY, D.B.A Harvard 1976. Temporary management systems. Application of a contingency theory to the creation of computer based information systems.

JOSE A. AMADOR, Ph.D. Florida 1977. Information formats and decision performance: An experimental investigation.

EMMANUEL APEL, Ph.D. Harvard 1977. Risk classes, systematic risk, and cost of capital in the chemical industry, 1957-68.

ARIE BARAN, Ph.D. California (Berkeley) 1977. Price level earnings: An empirical study of their time-series behavior and relationship to security price changes.

RODNEY G. BEDDOWS, D.B.A. Harvard 1976. National institutes of health: A study of the development and adaptation of administrative systems.

THOMAS V. BINGHAM, Ph.D. Cincinnati 1977. Economic models: The large, multiproduct firm and its organizational structure.

STEPHEN J. BROWN, Ph.D. Chicago 1976. Optimal portfolio choice under uncertainty: A Bayesian approach.

BERNARD R. CATRY, D.B.A. Harvard 1977. A comparison of the advertising and sales force role for industrial and consumer activities, respectively.

HOWARD A. CHERNICK, Ph.D. Pennsylvania 1976. The economics of bureaucratic behavior: An application to the allocation of federal project grants.

DONG SUNG CHO, D.B.A. Harvard 1976. International facility planning. Regarding the application of scientific approaches.

JESUS H. CHUA, Ph.D. Michigan 1976. Long-term corporate bonds as an investment medium.

HAROLD S. DIAMOND, JR , Ph D New York 1976 Pattern recognition and the detection of corporate failure.

STEVEN L. DIAMOND, D.B.A. Harvard 1977. The development of brand related attitudes, skills, and knowledge in children.

STEVEN X DOYLE, D.B.A. Harvard 1976. The motivation and compensation of field sales representatives.

YVES L DOZ, D.B.A. Harvard 1976. National policies and multinational management.

PAUL W. FARRIS, D B.A. Harvard 1976. Advertising intensity in consumer goods marketing: An analysis of variations in advertising-to-sales ratios.

NORMAN D. FAST, D.B.A Harvard 1977. The evolution of corporate new venture divisions

ROBERT L. GESKE, Ph.D California (Berkeley) 1977. The valuation of complex options.

DAVID M GOLDSMITH, Ph.D. Harvard 1977 The theory of optimal diversification.

ROBERT S. HARRIS, Ph.D. Princeton 1977. The impact of corporate mergers on acquiring firms: The 1954-1971 experience.

MICHAEL D. HAWKINS, Ph.D. California (Berkeley) 1977. Controlling transition rates in Markov manpower flow models.

WILLIAM A. HILLISON, Ph.D. Florida 1977. Empirical investigation of general purchasing power adjustments on earnings per share and the movement of securities prices.

R. BRUCE HUTTON, Ph.D. Florida 1977. Life cycle cost. The impact on the processing of new information for durable goods.

MARIANN JELINEK, D.B.A. Harvard 1977. Institutionalizing innovation.

JAI SEONG KANG, Ph.D. Texas A&M 1977. A study of intersegment allocation problems in financial re-

ports by segments of a diversified business enterprise.

BERNARD KITOUS, Ph.D. California (Berkeley) 1977. Interactive matrix displays and management reporting: A feasibility assessment.

HENRY W. LANE, D.B.A. Harvard 1977. Managing innovation: A comparative study of program level research and development units.

RUSSELL D. LANGER, Ph.D. California (Berkeley) 1977. Accounting as a variable in mergers.

LORMAN L. LUNSTEN, Ph.D. Michigan 1976. A model to improve the quality of the retail share of market forecast for banking offices.

GEORGE M. MCCABE, Ph.D. Pennsylvania 1975. The capital investment, earnings retention, and new debt behavior of large American corporations, 1967 1971.

JAMES M. MACLACHLAN, Ph.D. California (Berkeley) 1977. Response latencies and the prediction of consumer purchase probabilities.

JAMES R MENSCHING, Ph.D. Chicago 1976. Forward algorithms and planning horizons for deterministic inventory problems.

JUAN P. MONTERMOSO, D.B.A. Harvard 1977. The integrative role of information systems for agribusiness.

PHILIP D. NATHANSON, Ph.D. Michigan 1977 An empirical investigation of the relationship between common stock prices and consumer sentiment.

VIVIAN NOSSITER, D.B.A. Harvard 1976 The coordination of interdependent units. An empirical field study.

GEORGE PAPADATOS, Ph.D. Virginia Polytechnic Institute 1977. The bureaucratic theory of synthetic advertising.

DAVID L. PARRY, Ph.D. Southern California 1977 An investigation of Engel curves for various store types.

JOHN A. QUELCH, D B.A Harvard 1977 Measurement and implications of the relative importance of product attributes The nutrition factor in breakfast cereal brand choice behavior

ROBERT H. ROCK, D.B.A. Harvard 1976. The chief executive officer managing the human resources of the large diversified industrial company.

PETER C. RYAN, Ph D Massachusetts 1977. The effect on portfolio performance of certain call option strategies.

MICHAEL D. SHERMAN, Ph.D. Purdue 1977 A theoretic-simulation analysis of the role of hierarchical structures in bureaucratic organization: Special case of the firm.

SAUMITRA SIRCAR, D.B.A. Harvard 1976. The management of data for natural resources.

WILLIAM J. SMITH, Ph.D. Colorado 1976. Sensitivity of capital budgeting.

ROGER A. STRANG, D.B.A. Harvard 1977. Determining promotional strategy: An investigation of the advertising/sales promotion allocation decision process.

CHAM-KAU TAM, Ph.D. Toronto 1976. An econometric study of Canadian capital formation by industry.

MARYAM TASHAKORI, D.B.A. Harvard 1977. Management succession: from the owner-founder to the professional president.

PIERO TELESIO, D.B.A. Harvard 1977. Licensing policy in multinational enterprises.

CHARLES WALDMAN, D.B.A Harvard 1977. International mass retailing: A clinical analysis of current decision-making processes in selected European ventures.

JEROLD B. WARNER, Ph.D. Chicago 1976. Bankruptcy costs, absolute priority, and the pricing of risky debt claims.

WILLIAM WATERS, Ph.D. Chicago 1976 The effect of employer pension plan membership on the level of household wealth.

BIRGER WERNERFELT, D.B.A. Harvard 1977 An information based theory of microeconomics and its consequences for corporate strategy.

LARRY WYNANT, D.B.A. Harvard 1977. A study of financial strategies for large-scale mining ventures

EDWARD E. YARDENI, Ph.D. Yale 1976. A portfolio-balance approach to corporate finance.

DAVID W. YOUNG, D.B.A Harvard 1977. Organizational design and adaptation in social service agencies. An administrative systems perspective.

Industrial Organization and Public Policy; including Economics of Technological Change, and Industry Studies

TITUS O ADEBOYE, D.B.A Harvard 1977. International transfer of technology: A comparative study of differences in innovative behavior

JAMES H ALLEMAN, Ph.D. Colorado 1976 The pricing of local telephone service: A comparison of one- and two-part local telephone prices

ERNEST ANKRIM, Ph.D. Oregon 1976. Elasticities, biases and firm pricing strategy: An empirical and theoretical examination of cable television demand models.

CRAIG F. ANSLEY, Ph.D. Michigan 1977. Statistical analysis and econometric modeling of loss reserves in automobile liability insurance.

ALFRED B. ARTERBURN, Ph.D. Michigan 1977. Risk and the rate of return to capital.

CHRISTINE ASSIS, Ph.D. Johns Hopkins 1977. A mixed-integer programming model for the Brazilian cement industry

SAMUEL L. BAKER, Ph.D. Harvard 1977. Medical licensing laws: An early liberal reform.

HOWARD S. BANTON, Ph.D. Georgia State 1976. An empirical analysis of the spatial determinants of the demand for bank services.

EDWARD BEAUVAIS, Ph.D. Virginia Polytechnic Institute 1977. An interventionist model of public utility regulation.

GAIL R. BLATTENBERGER, Ph.D. Michigan 1977. Block rate pricing and the residential demand for electricity.

WILLIAM H. BONCHER, Ph.D. Indiana 1976. Innovation and technical adaptation in the Russian economy: The growth in unit power of the Russian mainline freight locomotive.

SANDFORD B. BORINS, Ph.D. Harvard 1976. The economics of airport planning: The case of Toronto.

PIETER BRAKEL, Ph.D. Toronto 1977. Production relationships, market structure, firm behavior, and the supply of exports: An application to the Canadian forest product industries.

JOHN R. CARTER, Ph D. Cornell 1976 An econometric study of the determinants and effects of corporate diversification

VLADIMIRO CATTO, Ph.D. Virginia 1977. An empirical determination of the effects of market power on performance

DAVID C CAVANDER, Ph.D. Iowa 1976 The *U S zinc* industry: An econometric study.

DAVID M. CHEREB, Ph.D. Southern California 1977 Endogenous technological change.

WILLIAM WING FAI CHOA, Ph.D. Stanford 1976 The relevance of market structure to technological progress: A case study of the chemical industry.

NATHANIEL B. CLARKE, Ph.D. Georgetown 1977. A theoretical evaluation of the sunk cost philosophy of public utility rate making.

JOSEPH CORDES, Ph.D. Wisconsin (Madison) 1977 The economics of "actual" compensation A theoretical and empirical perspective.

RONALD COTTERILL, Ph D. Wisconsin (Madison) 1977. Market structure, performance, and market restructuring in food retailing.

WILLIAM CROWELL, Ph D. New York 1977 An analysis of transit finance and pricing options

HUGH B. DAVIES, Ph.D. Pennsylvania 1975. Exit, voice, and the British national health service.

JAMES W. DOANE, Ph.D. Tufts 1977. The empirical basis for neoclassical production theory A study of capital-fuel substitution in electricity generation.

JEFFREY T. DOUTT, Ph.D. California (Berkeley) 1976. Productivity in fast-food retailing.

WILLIAM M. DUNN, Ph D. Florida 1976. The development, design and testing of a performance-measurement model for a not-for-profit organization.

DONNIE L. EVERETTE, D B.A. Harvard 1977. The Federal Trade Commission and consumer representation.

YEHUDA FRIEDENBERG, Ph.D. Chicago 1976. Price signalling and equilibrium price distributions.

CYNTHIA B. GLASSMAN, Ph.D. Pennsylvania 1975. The effects of mergers on price and service performance of Pennsylvania banks.

OLIVER GOLDSMITH, Ph.D. Wisconsin (Madison) 1976. Regulation of a nonrenewable resource: The case of natural gas.

ELLEN W. GOLDSTEIN, Ph.D. Brown 1977. An analysis of tie-in contracts.

DAVID R. GRAHAM, Ph.D. California (Los Angeles) 1976. A test of the Averch-Johnson model of regulation using electric utility data.

TIMOTHY S. GREENING, Ph.D Harvard 1976. Oil wells, pipelines, refineries, and gas stations: A study of vertical integration

JOSEPH A. GRIBBIN, Ph.D. Catholic 1977. An analysis of cyclical fluctuations in commercial shipbuilding in the United States, 1945-75.

JOHN K GROW, Ph.D Cincinnati 1977. Market influences on technological change.

ARTHUR C. GRUEN, Ph.D Tufts 1977 The efficiency and equity of a congestion toll on automobile traffic in the Boston metropolitan area

DAVID S. HABR, Ph D Washington State 1976. The returns to advertising. An examination of the relationship between advertising and liquor sales in the state of Washington

LARRY J HAVERKAMP, Ph.D. Pennsylvania 1977. The measurement and explanation of monopoly returns.

JOHN F. HENNIGAN, JR, Ph.D West Virginia 1976 Economic differences between public and private ownership. The water utility case.

HARLOW H. HIGINBOTHAM, Ph.D Chicago 1976. The demand for hedging in grain futures markets

MARK HIRSCHEY, Ph D Wisconsin (Madison) 1977. An incentive contract methodology for subsidized railroad freight services.

DANIEL R HOYT, Ph.D. Nebraska 1976 Implications of direct and indirect interlocking directorships for effective competition.

HAMID JARAIEDI, Ph D Georgia 1976 A disaggregated analysis of the demand for gasoline in the United States

JAMES KEDDIE, Ph.D. Harvard 1976. Adoptions of production technique in Indonesian industry.

PAUL KOBRIN, Ph.D New York 1976. Vertical integration and efficiency in the petroleum industry.

JOSEPH W. LEVEDAHL, Ph D. Chicago 1976. The effects of household's permanent and transitory income on both their time of automobile purchases and their net outlay for automobile services for 1968, with predictions for 1969.

BARRY LITMAN, Ph.D. Michigan State 1976 Vertical integration and performance of the television networks.

JAMES T. LOW, Ph.D. Michigan 1977. A model for evaluation of alternative means of reducing airport system congestion considering cost, delay, and noise exposure criteria.

DENNIS L. MCNEILL, Ph.D. Florida 1977. An investigation of issues surrounding the presentation of energy cost information for consumer durables.

DOROTHY M. MERCER, Ph.D. Oklahoma State 1977. An economic analysis of the technical, allocative, and equity effects of financing municipal government from revenue over cost earned by the municipal electric utility: A case study of Stillwater, Oklahoma.

GUY C. Z. MHONE, Ph.D. Syracuse 1977. Market in the copper industry in Zambia, 1929 69

CAROL D. NORLING, Ph.D. Minnesota 1977. Demand for gasoline.

DENNIS C. O'NEILL, Ph.D. Cincinnati 1976. An optimal approach to the theory of conglomerate growth.

ALAN OSMAN, Ph.D. Ohio State 1977. Benefits and costs of specific manufacturing firm locations: Case study of a five-county region in southeast Ohio.

KEITH C. PARK, Ph.D. Catholic 1976. An economic analysis of trends in Japanese textile industry, 1953-1973.

LAURENCE T. PHILLIPS, Ph.D. Utah 1976. Industrial symmetry and its influence in oligopolistic markets.

MICHAEL J. PIETTE, Ph.D. Florida State 1977. The *U.S.* petroleum pipeline industry: A study of vertical integration.

CHUNG LAM POON, Ph.D. Western Ontario 1976. Urban railway relocation: An economic evaluation.

P. SOMESWARA RAO, Ph.D. Queen's 1977. Forecasting the demand for railway freight transportation services: An integration of econometric modelling and input-output analysis.

EUGENE F. RASMUSSEN, Ph.D. Nebraska 1976. Federal Power Commission regulation of wholesale electric sales: A case study of policy initiation.

DOUGLAS K. REECE, Ph.D. California (Berkeley) 1977. Leasing offshore oil: An analysis of alternative information and bidding systems.

FRIEDA C. REITMAN, Ph.D. New School 1977. The measurement and analysis of demand: Elasticities of residential vertical telephone service.

PATRICIA M. RUDOLPH, Ph.D. North Carolina (Chapel Hill) 1976. Industry, size and cyclical differences in the balance sheet structure of manufacturers· 1964–74.

LESTER F. SAFT. Ph.D. California (Los Angeles) 1976. Franchising, the requirements contract, and the Chicken Delight decision.

SHELTON S. SCHMIDT, Ph.D. Virginia 1977. An empirical evaluation of the full-cost theory of price determination.

GARY E. SELLERS, Ph.D. Cincinnati 1977. Regulating community antenna television by local franchise: The Ohio experience.

MEREDITH SLOBOD, Ph.D. California (Los Angeles) 1977. Industrial demand for electricity: Analysis of pooled data of a single utility.

RODNEY T. SMITH, Ph.D. Chicago 1976. The legal and illegal markets for taxed goods and taxation policy: Pure theory and an application to state government taxation of distilled spirits.

CHARLES W. SMITHSON, Ph.D. Tulane 1976. The degree of regulation and the monopoly firm.

FRANCISCO A. S. DE SOUZA, Ph.D. Vanderbilt 1977. An economic analysis of the greater Sao Paulo fluid milk market.

MARION STEWART, Ph.D. Columbia 1976. Essays on market structure, risk, and optimal technological progress.

MICHAEL A. STOLLER, Ph.D. Washington (St. Louis) 1977. The economics of UHF television: Effects of government policy.

SCOT A. STRADLEY, Ph.D. Utah 1977. Human resource implications of the production process in underground bituminous extraction, especially for Utah.

PETER STRATTON, Ph.D. Northern Illinois 1976. Public libraries: Their structure, use, and cost of service provision.

VATCHAREEYA THOSANGUAN, Ph.D. Duke 1977. Net rates of entry and their determinants in *U S* manufacturing industries, 1954–72.

JOHN W. TWAY, Ph.D. Oklahoma 1977. The economics of multihospital organizations and efficiency of rural hospitals.

MICHAEL C. VARNER, Ph.D. Cornell 1976. Advertising and conglomeration in the food processing industries.

W. KIP VISCUSI, Ph.D. Harvard 1976. Employment hazards An investigation of market performance.

WILLIAM P. WELCH, Ph.D. Colorado 1977. Resources in political campaigns: A unified theory.

JEFFREY R. WILLIAMS, Ph.D. Michigan 1977. Business risk and financial policy in the electric utility industry.

JOHN M. WILLS, Ph.D. Washington 1976. Technical change in industrial use of energy outputs.

MICHAEL A. Zimmer, Ph.D. Tennessee 1977. Cost of capital and regulatory effectiveness in the electric utility industry: An empirical analysis.

Agriculture and Natural Resources

KAMEL SELMAN AL-ANI, Ph.D. Pennsylvania State 1977. The relationship of water quality to residential property value.

ROY L. ALLEN, Ph.D. Colorado State 1976. The petroleum industry: Public or private?

JERRY W. ANDERSON, Ph.D. Cincinnati 1976. Air pollution's costs and benefits: A present value approach to the Cincinnati air basin.

JAMES H. BASKETT, Ph.D. Arizona 1976. Economic impacts of U.S. Forest Service policies on local communities: An interindustry analysis of the Salt-Verde Basin, Arizona.

JAMES G. BEIERLEIN, Ph.D Purdue 1977. Optimizing the capital structure of farmer cooperatives using a member oriented analysis.

YOAV BENARI, Ph.D. Princeton 1977. Depletable resource pricing and output strategies in light of a possible future substitute.

MARTIN J. BLAKE, Ph.D. Missouri (Columbia) 1976. Optimal storage policies for wheat.

MOHSEN BOLOORFOROOSH, Ph.D. Iowa State 1977. Demand estimation of meat in Iran

LARS G. BRINK, Ph.D. Purdue 1976. Plans, decisions, and results: An evaluation of a procedure for farm planning under risk.

FRANK A CAMM, JR. Ph D. Chicago 1976 Political economy and agricultural marketing orders.

MARIO M. CARILLO-HUERTA, Ph D Vanderbilt 1976. Agricultural credit, insurance, and the adoption of innovations by Mexican farmers

SEOK-JUNG CHANG, Ph.D. North Carolina State 1976 An economic analysis of the adoption of new wheat varieties under uncertainty in Ferozepur District, India.

KENNETH C. CLAYTON, Ph.D. Purdue 1976 A planning model for regional solid waste management systems.

JOHN M. CURRIE, Ph.D. California (Berkeley) 1976. Property rights and agricultural land An examination of the economics of agricultural land tenure in England and Wales

GARY A DAVIS, Ph D. Michigan State 1977. A simulation study of decision strategies for hog processing plants constrained by energy supply regulations.

WILLIAM H. DESVOUSGES, Ph.D. Florida State 1977. Competition in outer continental shelf lease sales, 1954 74

DONALD E ECKERMAN, Ph D. California (Berkeley) 1976. The effects of reservation systems, income and season on use of California campgrounds.

DOEKE C. FABER, Ph D. Iowa State 1976 Can agriculture feed the world: A world food analysis.

WAN FEE FEE, Ph.D. Wisconsin (Madison) 1976. Land resource utilization and land use controls in West Malaysia, 1957 74.

JAMES C. FINK, JR., Ph.D. Pennsylvania State 1977 An economic model for estimating work trip distributions between minor civil divisions in rural areas.

DEREK FORD, Ph.D. Pennsylvania 1977. Coffee supply, trade, and demand: An econometric analysis of the world market, 1930-69.

FRANKLIN G. FOX, JR., Ph.D. Washington 1977. Bureau of Indian Affairs timber sales policies: Economic impact on the Flathead Indian Reservation.

DAVID FRESHWATER, Ph.D. Michigan State 1977. The linkages between individual use and public management of flood plains.

WINSHIP C. FULLER, Ph.D. Tufts 1977. The transformation of Chinese agriculture, 1949-74.

ATOS F. GRAWUNDER, Ph.D. Wisconsin (Madison) 1976. The southern Brazil agricultural sector: The income problem.

JOHN J. J. GREGOR, Ph.D. Pennsylvania State 1976. Air quality and intraurban mortality: An economic analysis of the costs of pollution-induced mortality.

JOEL GUTTMAN, Ph.D. Chicago 1976. The demand for publicly financed agricultural research: An application of a theory of collective action

WILLIAM F. HAMPEL, JR., Ph D Iowa State 1976. A maximum bid-price model to evaluate factors influencing farmland values.

DALE L HULL, Ph D McGill 1976. The programming evaluation of spatial and intertemporal allocation policies with respect to interior provincial Crown forest land in British Columbia.

ISAIAH I IHIMODU, Ph.D. Pennsylvania 1977. Impact of government taxation and control of the marketing of five major agricultural exports of Nigeria.

DAVID C ILLIG, Ph D. California (Santa Barbara) 1977 Prices and buffer stocks: An analysis of the impact of a futures contract on decision making in the *U S* softwood plywood industry.

DOYLE JEON, Ph.D. Kansas State 1976 World wheat economy Development of dynamic analysis of an imperfect market.

JAGDEESH C KALLA, Ph.D. Ohio State 1977. Saving investment behavior of farm families Udaipur District, Rajasthan, India.

THOMAS J. KERRIGAN, Ph.D Fordham 1977. Influences upon the future international demand and supply for coke.

KURT K. KLEIN, Ph.D. Purdue 1976. An investigation into the evaluation of applied research projects in agriculture.

MAHLON G. LANG, Ph.D. Michigan State 1977 Collective bargaining between producers and handlers of fruits and vegetables for processing: Setting laws, alternative rules, and selected consequences.

DAVID L. LANGEMEIER, Ph.D. Nebraska 1976. A microeconomic analysis of the effects of hail suppression.

KENNETH L. LEATHERS, Ph.D. Colorado State 1976. The economics of controlling saline irrigation return flows. A case study of the Grand Valley, Colorado.

LINDA K. LEE, Ph D. Iowa State 1977. The conversion of agricultural land to urban use: An examination of the land use process in Urbandale, Iowa, 1950–1974.

WILLIAM G. LESHER, Ph.D. Cornell 1976. An analysis of northeastern rural land use policies.

JOHN MCCALLUM, Ph.D. McGill 1977. Agriculture and economic development in Quebec and Ontario to 1870.

ROBERT N. MCRAE, Ph.D. British Columbia 1977. A quantitative analysis of some policy alternatives affecting Canadian natural gas and crude oil demand and supply.

Sonia MALTEZOU, Ph.D. New York 1977. The economic determinants of waste oil recycling.

CRAIG S. MARXSEN, Ph.D. Georgia State 1976. The influence of interest and inflation on fuels supply.

ABRAHAM MELAMED, Ph.D. California (Berkeley) 1977. The citrus marketing board of Israel and the auction demand for weekly sales.

DONALD O. MITCHELL, Ph.D. Iowa State 1976. The level and variability of agricultural exports.

EDGARDO R. MOSCARDI, Ph.D. California (Berkeley) (1977). A behavioral model for decision under risk among small-holding farmers.

D. DAVID MOYER, Ph.D. Wisconsin (Madison) 1977. An economic analysis of the land title records system.

JOHN A. MURPHY, Ph.D. California (Berkeley) 1976. An economic study of the Irish cattle industry.

ISMAIL H. OZSABUNCUOGLU, Ph.D. Arizona 1977. Economic impacts of alternative irrigation systems under increasing irrigation water costs in southeastern Arizona.

DAVID H. PETERSON, Ph.D. Wisconsin (Madison) 1977. Efficiency and equity in pricing local public services: The case of wastewater.

TODD E. PETZEL, Ph.D. Chicago 1976. Education and the dynamics of supply.

RULON D. POPE, Ph.D. California (Berkeley) 1976. An analysis of factors affecting farm diversification with special reference to San Joaquin Valley crop farms.

ALLEN M. PRINDLE, Ph.D. Pennsylvania State 1977. A static equilibrium model of the *U.S.* feed-livestock sector.

SHIRLEY A. PRYOR, Ph.D. Michigan State 1977. The implications of recent economic and policy changes on retail and farm level demand for food commodities.

CHANDRASHEKHAR G. RANADE, Ph.D. Cornell 1976. Distribution of benefits from new agricultural technologist: A study at farm level.

CAROLEE SANTMYER, Ph.D. Washington State 1976. An economic analysis of alternative long-term management programs for sawlogs, beef, elk, and deer on a small portion of the St. Joe National Forest.

KADIR D. SATIROGLU, Ph.D. Colorado State 1976. Analysis of agricultural production in the United States, 1950–74.

KENNETH J. SAULTER, Ph.D. California (Santa Barbara) 1976. Energy demand in *U.S.* manufacturing: A cross-section study for 1958, 1962, and 1971.

ROBERT R. SCHNEIDER, Ph.D. Wisconsin (Madison) 1976. Diffuse agricultural pollution: The economic analysis of alternative controls.

ROGER E. SCHNEIDER, Ph.D. Missouri (Columbia) 1976. An evaluation of disciplinary analysis of entry and exit in commercial agriculture.

BRYAN W. SCHURLE, Ph.D. Ohio State 1977. An analysis of the return-risk tradeoffs associated with tomato production in northwestern Ohio.

JOSEF SCHWALBE, Ph.D. Washington 1977. Investigation of factors effecting demand for national parks.

BRENT W. SPAULDING, Ph.D. Iowa State 1976. Economic impacts on agriculture in the north central region of the United States under alternative land use policies: A modeling approach.

THOMAS H. SPREEN, Ph.D. Purdue 1977. An application of capital theory in a recursive linear programming model of the cattle subsector of Guyana.

STEPHEN STALOFF, Ph.D. Oregon 1977. A stock-flow analysis of copper markets.

VASANT SUKHATME, Ph.D. Chicago 1977. The utilization of high-yielding rice and wheat varieties in India: An economic assessment.

RICHARD W. THORESON, Ph.D. Harvard 1976. Resource depletion and federal price regulation of natural gas.

W. BRUCE TRAILL, Ph.D. Cornell 1976. Incorporation of risk variables in econometric supply response analysis.

JAMES N. TRAPP, Ph.D. Michigan State 1976. An econometric simulation model of the *U.S.* agricultural sector.

MOHAMMAD USMAN, Ph.D. Colorado State 1976. Short-run beef demand and supply in the United States.

ASHTON I. VERAMALLAY, Ph.D. Iowa State 1976. An evaluation of water resource developments in Guyana: With application to selected drainage and irrigation projects.

GARY F. VOCKE, Ph.D. Iowa State 1977. Economic and environmental impacts on *U.S.* agriculture from insecticide, fertilizer, soil loss, and animal waste regulatory policies.

GREGORY K. WHITE, Ph.D. Washington State 1976. The impact of the pine bark beetle on recreational values.

JAMES K. WHITTAKER, Ph.D. North Carolina State 1977. A comparison of alternative formulations of econometric crop supply response models.

Manpower, Labor, and Population; including Trade Unions and Collective Bargaining

RICHARD D. ANSON, Ph.D. Georgetown 1977. An economic evaluation of the monetary and non-

monetary private effects of migration: The case of Monterrey, Mexico.

ANDRE ARCHER, Ph.D. City (New York) 1976. Internal migration and traditional agriculture: The case of Puerto Rico.

JOE G. BAKER, Ph.D. Utah 1977. Labor allocation and western energy development.

HIROYA BAMBA, Ph.D. Massachusetts 1977. Relationship between the structure of the labor market and demand for education by young black and white males.

MARK A. BEBENSEE, Ph.D. Duke 1977. Income distribution effects of increased labor force participation by married women.

BRIAN E. BECKER, Ph.D. Wisconsin (Madison) 1977. The impact of unions on labor costs in hospitals: A three-state study.

CHARLES L. BETSEY, Ph.D. Michigan 1976. Work experience and earnings of inner-city blacks and whites

MATTHEW BLACK, Ph.D. Michigan 1976. The determinants and implication of job search and quit activity.

RICHARD N. BLOCK, Ph.D. Cornell 1977. The impact of union-negotiated job security provisions on labor turnover and labor mobility

RICHARD BURKHAUSER, Ph.D. Chicago 1976. An early pension decision and its effect on exit from the labor market.

MICKEY L. BURNIM, Ph.D. Wisconsin (Madison) 1976. Comparative rates of return to black college graduates: The black versus the nonblack schools.

GERALDO S. DE CAMARGO BARROS, Ph.D. North Carolina State 1976. Asking wages, market wages, and the off-farm labor supply by farm operators.

JOHN F. COGAN, Ph.D. California (Los Angeles) 1976. Reservation wages, labor force participation rates, and hours of work of married women.

JOSEPH A. COLOSI, Ph.D. Wisconsin (Madison) 1977. Simulated family planning in less developed countries: A disaggregated approach.

LESLIE G. DENEND, Ph.D. Stanford 1977. An economic investigation of the expected returns across occupations in the nonsupervisory *U.S* labor force.

FREDERICK W. DERRICK, Ph.D. North Carolina State 1976. The work decision of college students.

HUGO C. DIAZ-ETCHEVEHERE, Ph.D. Wisconsin (Milwaukee) 1977. An analysis of earnings differentials between Anglos, Mexican-Americans, and blacks in five southwestern states.

E. JOHN DRIFFILL, Ph.D. Princeton 1977. Essays in labor supply and human capital accumulation.

PETER J. EATON, Ph.D. Florida 1976. The in-migration of unskilled labor to large urban centers in the Brazilian northeast.

WILLIAM H. EPSTEIN, Ph.D. Harvard 1977. Schooling and occupation decisions.

HABIB T. FATOO, Ph.D. Michigan State 1977. Macroeconomic analysis of output, employment, and migration in Sierra Leone.

ROBERT M. FEINBERG, Ph.D. Virginia 1976. Theoretical implications and empirical tests of the job search theory.

PETER C. FELSTED, Ph.D. Vanderbilt 1976. The productive and distributive impact of educational investment in Peru: A sector approach.

LAURENCE R. FISHER, Ph.D. Connecticut 1977. The adequacy of workmen's compensation death benefits in Connecticut.

CHRISTOPHER R. P. FRASER, Ph.D. Wisconsin (Madison) 1977. The impact of collective bargaining on job satisfaction.

LUCIAN B. GATEWOOD, Ph.D. Wisconsin (Madison) 1977. A comprehensive evaluation of Recruitment and Training Program Inc.'s apprenticeship outreach program for the construction trades: An analysis of program performance and impact.

RONALD T. GILLASPY, Ph.D. Pennsylvania State 1976. Dynamics of sex mix and desired family size.

RICHARD E. GINNOLD, Ph.D. Wisconsin (Madison) 1976. A follow-up of permanent disability under Wisconsin worker's compensation.

WILLIAM B. GOODWIN, Ph.D. Purdue 1976. Relative wages, wage expectations, and controls.

GARY A HALL, Ph.D. Tennessee (Knoxville) 1976. American wage arbitration, 1950–74.

STEPHEN K. HAPPEL, Ph.D. Duke 1976. The economics of the timing and spacing of births

DANIEL H. HILL, Ph.D. Michigan 1977. A dynamic analysis of the labor force participation of married women.

BARRY T. HIRSCH, Ph.D. Virginia 1977. The level and distribution of earnings across *U.S* labor markets. A microeconomic approach.

HOWARD R. HOGAN, Ph.D. Princeton 1976. Age patterns of infant mortality.

KEVIN HOLLENBECK, Ph.D. Wisconsin (Madison) 1976. The formulation and application of a model of the employment and earnings incidence of public policy to the analysis of the regulation of air pollution.

JOHN W. HOLMES, Ph.D. Michigan 1976. A modified Markov process of income dynamics

GILBERT G. JOHNSON, Ph.D. Yale 1977. Nonunion wage changes in Canada.

STEPHEN D. KANYR, Ph.D. North Carolina State 1976. The allocation of human capital over the working life cycle: The case of self-employed individuals.

NICHOLAS M. KIEFER, Ph.D. Princeton 1976. Econometric essays in labor economics.

CHRISTOPHER KING, Ph.D. Michigan State 1976. The unemployment impact of the Vietnam War.

DAVID P. KING, Ph.D. Tulane 1976. Speculative adjustment and market efficiency: A generalized model of labor migration.

MICHAEL L. KLIMA, Ph.D. Washington State 1977. An analysis of the quit rate for the manufacturing sector and major industry groups in manufacturing, 1950–70: A two-sector model.

JAMES E. KOCHER, Ph.D. Michigan State 1976. A microeconomic analysis of the determinants of human fertility in rural northeastern Tanzania.

MICHAEL H. KUPILIK, Ph.D. Colorado 1976. Age structure of the population and the distribution of life cycle income.

MHAMED LAKHOUA, Ph.D. Michigan State 1976. Cost-benefit analysis of exporting workers: The Tunisian case.

PAUL C. LANGLEY, Ph.D. Queen's 1977 Migration, earnings differentials, and past migratory experience: Canada and the United States.

BERNARD F. LENTZ, Ph.D. Yale 1976. Public sector wage determination: A democratic theory of economics.

SAMUEL S. LIEBERMAN, Ph.D. Harvard 1976. An economic approach to differential demographic behavior in Turkey

PAK WAI LIU, Ph.D. Stanford 1976. The economics of work effort and alternative wage payment systems.

WILLIAM F. MALONEY, Ph.D. Michigan 1976 Productivity bargaining: A study in contract construction.

FRANCISCO M. MOCHON, Ph.D. Indiana 1976 Human differences and human and physical capital flows A case study of Spain

ROBERT L. MOORE, Ph.D. Harvard 1977 Job search, job changing, and unemployment among young men, financial aid to college students, and the CSS need analysis system

CARL A. MOSK, Ph D Harvard 1976 Demographic transition in Japan, 1920 60

ERIC R. NELSON, Ph.D. Yale 1976 African rural-urban migration: Economic choice theory and Kinshasa evidence.

HAROLD D. NORDIN, Ph.D Illinois 1976 Comparative wage-income distribution through population shifts for Great Britain, Sweden, and the United States.

GADIS NOWELL, Ph.D. Chicago 1976 A study of black students and their experiences in the "careers for blacks in management program "

MAHMOOD A.K. OMARZAI, Ph.D. Indiana 1977. A comparative economic analysis of the size distribution of incomes of white, black, and Spanish-American ethnic groups in the United States.

JOHN OSTBO, Ph.D. Utah 1976. Enrollment in the disability pension program of Norway: An application of the work-leisure theory of labor supply.

EPICTETUS E. P. PATALINGHUG, Ph.D. Hawaii 1977. Education, productivity, and economic growth: A two-sector analysis of alternative measures of qualitative change.

MICHAEL B. PERCY, Ph.D. Queen's 1977. Migration flows during the decade of the wheat boom in Canada, 1900-10. A neoclassical analysis.

HARLAN D. PLATT, Ph.D. Michigan 1976. A simultaneous equations model of wage determination under collective bargaining.

MARSHALL I POMER, Ph.D. Harvard 1976. Occupations of fathers and sons: An application of the segmentation perspective to the OCG data on intergenerational mobility.

ALLEN M. PONAK, Ph.D. Wisconsin (Madison) 1977. Registered nurses and collective bargaining: An analysis of job related goals.

GARY R RAY, Ph.D. California (Santa Barbara) 1976. Production in the family economy.

MOHAMED S REDJEB, Ph.D Stanford 1977. A longitudinal study of the determinants of success in the Tunisian labor market, socioeconomic background, ability, and schooling

BARBARA J REDMAN, Ph D. Iowa State 1976 Determinants of parental investments in children's human capital formation via higher education

JEFFREY A. REED, Ph D Louisiana State 1976 Frictional, structural, and cyclical factors in Louisiana unemployment

CORDELIA REIMERS, Ph.D Columbia 1977. Factors affecting the retirement decision of American men

RONALD RIZZUTO, Ph D. New York 1977. The economic impact of school quality on earnings differentials and income distribution

MARC RUBIN, Ph D Pennsylvania 1977 Soviet manpower planning Decentralized action within a centralized framework

WAYNE F. RUHTER, Ph.D California (Los Angeles) 1976. Human capital development and parental incentives

PAUL RYAN, Ph D Harvard 1977. Job training.

FARUK G. SABET, Ph D. New School 1977. The political economy of the determinants of population change.

HENRY SAFFER, Ph D. City (New York) 1977 A structural model of wage differentials between men and women in professional employment.

MARCUS G SANDVER, Ph.D. Wisconsin (Madison) 1976. The trends and determinants of salary of major officials of local and national unions, 1962-73

DAVID E. SISK, Ph.D. California (Los Angeles) 1976. Ph.D. production in public and private universities

STANLEY K. SMITH, Ph.D. Michigan 1976. Women's work and fertility in Mexico City.

RICHARD S. SNYDER, Ph.D. Indiana 1976. The effect of ADC work incentives on labor force participation of ADC mothers in Indiana.

LAWRENCE M. SPIZMAN, Ph.D. State University of New York (Albany) 1977. Public sector employment decisions: An econometric analysis.

JAMES B. STEWART, Ph.D. Notre Dame 1976. Relative levels of economic welfare of the black population in the United States and South Africa.

RICHARD W. STRATTON, Ph.D. Connecticut 1977. Local labor market structure: Its effect on wages.

LADDISLAV J. SULA, Ph.D. Georgia State 1976. Time allocation in the home: Intergenerational wealth effects of the negative income tax.

B. KATHERINE SWARTZ, Ph.D. Wisconsin (Madison) 1976. Screening in the labor market: A case study.

BANKEY B. TANDON, Ph.D. Queen's 1977. An empirical analysis of earnings of foreign-born and native-born Canadians.

JAMES R. TARR, Ph.D. Florida State 1977. Occupational migration flows into and from Florida: 1965-70.

PAUL A. TATSCH, Ph.D. State University of New York (Binghamton) 1977 A study of household determinants of migration.

ABULKASEM O TOBOLI, Ph D. Oklahoma State 1976. An economic analysis of internal migration in the Libyan Arab Republic.

MICHIHIDE D. TOTOKI, Ph.D. Utah 1977. An evaluation of Japanese manpower policy, 1955-70.

SHUNICHIRO UMETANI Ph.D. Wisconsin (Madison) 1977. The labor market and the rate of return to higher education in postwar Japan, 1954-73.

JOHN T. WARNER, Ph.D. North Carolina State 1976. Empirical tests of job search hypotheses. The generation and acceptance of job offers.

ANDREW M. WEISS, Ph.D. Stanford 1977. A theory of limited labor markets.

JON D. WELLAND, Ph.D. Minnesota 1976. An investigation of the role of abilities in the individual earnings function.

LOUIS L. WILDE, Ph.D Rochester 1977. Labor market equilibrium under imperfect information.

ROBERT M. WILSON, Ph.D. Northwestern 1977 Occupational choice and the decision to invest in formal education.

ROY A. WYSCARVER, Ph.D. Clark 1977. Inequality in the primary distribution of labor income A theoretical and empirical analysis.

EDWARD G. YOUNG, Ph.D. Washington State 1977. The use of evaluation as a performance incentive in a decentralized system of employment and training program planning.

Welfare Programs; Consumer Economics; Urban and Regional Economics

VINOD AGARWAL, Ph.D. California (Santa Barbara) 1977. Foreign students' demand for *U.S.* higher education.

JOHN ANDERSON, Ph.D. Claremont 1977. Determinants of house value: An application of ridge regression.

EMILY S. ANDREWS, Ph.D. Pennsylvania 1976. Earnings and sex: An empirical investigation of discrimination, 1940-51.

DAVID L. ARTHUR, Ph.D. California (Berkeley) 1977. An analysis of the changing decision-making roles of business and government in regional development: Related policy issues.

FRANK T. AYERS, Ph.D. Georgia State 1977. An empirical evaluation of the Von Thünen-Haig theory of land value determination: The case of I-64 and Kentucky.

STEVEN BALKIN, Ph.D. Wayne State 1976. Urban crime: An economist's perspective.

RICHARD J. BALLMAN, Ph.D. Iowa 1977. Intersectoral flows analysis: Application to the Quad-Cities economy and some tests of its explanatory powers.

MORRIS L. BARER, Ph.D. British Columbia 1977. A methodology for derivation of marginal costs of hospital cases, and application to estimation of cost savings from community health centers.

DAVID L. BARKLEY, Ph.D. Iowa State 1976. The factors contributing to industrial outmigration from rural areas of Iowa

KEMAL BAS, Ph.D. New School 1977. The role of education in reducing welfare expenditures in the long run.

BRUCE BENDER, Ph D Chicago 1976 The determinants of housing abandonment.

LARRY A BLAIN, Ph.D. British Columbia 1977. Regional cyclical behavior and sensitivity in Canada, 1919 73.

GLENN BLOMQUIST, Ph.D Chicago 1977. Value of life: Implications of automobile seat belt use.

ALAN J. BROKAW, Ph.D. Michigan 1976. An explanation of the preferences of suppliers for buyers.

WILLIAM A. N. BROWN, Ph.D. Manitoba 1977. The role of outdoor recreation in regional development: A study of Hecla Provincial Park.

BARRY BROWNSTEIN, Ph.D. Rutgers 1977. Income redistributive effects of public disability programs.

JAN K. BRUECKNER, Ph.D. Stanford 1976 An urban economic model with locally produced goods.

NORBERT W. BUDDE, Ph.D. Washington (St. Louis) 1976. Demand for hygienists. An intrafirm analysis.

JACEK J. CIANCIARA, Ph.D. Wisconsin (Milwaukee) 1977. A theory of spatial-dual economic growth.

ROBERT J. CUOMO, Ph.D. Boston College 1977. A model of the residential location decision: An attempt to incorporate and measure the importance of public goods in the decision-making process.

DAVID M. CURRIE, Ph.D. Southern California 1977. Interdisciplinary research and the university reward system.

JOHN J. DOYLE, JR., Ph D Clark 1976. The economic impact of institutions of higher learning upon the city of Worcester, Massachusetts and the surrounding region: A case study.

RICHARD F. DYE, Ph.D. Michigan 1976. Personal charitable contributions: Tax effects and other motives.

JOSEPH K. ECKERT, Ph.D. Tufts 1977. The effect of rent controls on assessment practices, differential incidence of taxation, and income adjustment mechanisms for rental housing in Brookline, Massachusetts.

RONALD C. FAAS, Ph.D. Michigan State 1977. The economics of budget constraints, transaction costs, boundaries and articulation of interest group demand in multicounty health planning agencies.

ROGER D. FELDMAN, Ph.D. Rochester 1977. The supply and demand for physicians' hospital and office visits.

DENNIS K. FUKUMOTO, Ph.D. Southern California 1977 Chinese and Japanese in California 1900 1920: A case study of the impact of discrimination

DANIEL L GEORGIANNA, Ph.D. Massachusetts 1977. Neighborhood effects and the structure of cities: A study of the relationship between variation in income and distance from the center of *U S* cities.

JOHN T. GILMORE, Ph.D. Colorado State 1976. Community development Theory and process

DIANE E. GOLD, Ph.D. New School 1977. Class structure and the economics of the housing problem

JOHN S. GREENLEES, Ph.D. California (Los Angeles) 1977. Discrimination against medicaid patients in nursing home admissions policies

ROY J. GROHS, Ph.D. Washington State 1977. The structure and determinants of county retail sales and service receipts and of county employment in Minnesota.

AARON S. GURWITZ, Ph.D. Stanford 1977. Local taxation and the dynamics of metropolitan property values.

ROBERT A. HANKIN, Ph.D. Northeastern 1976 The cost of providing restorative dentistry in an alternative delivery mode.

DAVID R. HENDERSON, Ph D. California (Los Angeles) 1976. The economics of safety legislation in underground coal mining.

MICHEL L. HISER, Ph.D. Cornell 1976 An interindustry analysis of Clinton County, New York.

DOUGLAS E. HOUGH, Ph.D. Wisconsin (Madison) 1976. The economics of the market for human blood

ALEXANDER W. JENKINS, Ph.D. Western Ontario 1977. A policy-oriented analysis of hospital costs in Ontario.

JOHN M. JOSEPH, JR., Ph.D Georgetown 1977 An econometric model of the state of Maine

DAVID KENNETT, Ph.D. Columbia 1977. A study of the distribution of a free public good: The case of crime and police services.

LAWRENCE KENNY, Ph.D. Chicago 1977. The demands for child quality and for educational inputs, the production of child quality, and related topics.

BODO KULJURGIS, Ph.D. Wayne State 1977. A regional econometric model of residential construction: An application to the Detroit metropolitan area.

DAVID LAH, Ph.D. Wisconsin (Madison) 1976. Causes and consequences of poverty and inequality in the inner city· An in-depth study.

CHARLES J. LATOS, Ph.D. Brown 1977. Determinants of intraregional differentials in nonwhite urban settlement patterns.

WILLIAM C. LEE, Ph.D. California (Santa Barbara) 1976. The demand for travel and the gasoline crisis.

WILLIAM E. LEGG, Ph.D. Florida State 1977. The theory and measurement of tenure choice in housing.

EUGENE LEWIT, Ph.D. City (New York) 1977. Experience with pregnancy, the demand for prenatal care, and the production of surviving infants

ROGER L. MCCLUNG, Ph.D Washington (St. Louis) 1976. Identification of an educational production function for diverse technologies

THOMAS G. MCGUIRE, Ph D. Yale 1976. Demand for physicians' services

EDWARD M. MCNERTNEY, Ph D Massachusetts 1977 An analysis of the effect of taxation on the location of industrial employment

BRUCE D MANN, Ph D. Indiana 1976. The influence of educational services on housing values Indianapolis, 1971 74.

JONATHAN H. MARK, Ph.D Washington (St Louis) 1977. Externalities and housing values

ROGER E. MIENERS, Ph.D Virginia Polytechnic Institute 1977 The economics of victim compensation

CASIMIRO MIRANDA, Ph.D Wayne State 1977. The regional pattern of industrial location A study of the Philippine space economy.

NICHOLAS L. NICHOLAS, Ph.D. Florida 1977. Identification of the innovative buyer in the retirement market.

JAIME NINO GAJARDO, Ph.D Stanford 1977. Food consumption patterns in the United States: An application of the Houthakker and Taylor dynamic demand model.

JYRKI T. PENTTINEN, Ph.D. Michigan 1976 The role of price in perception of product quality.

STEVEN G. PHILLIPS, Ph.D. Georgetown 1977. Toward a theory of optimal city size.

ROBERT POLLARD, Ph.D. Chicago 1977. Topographic amenities and building height in an urban housing model.

HENRY J. RAIMONDO, Ph.D. Wisconsin (Madison) 1976. The tax and expenditure implications of al-

ternate finance schemes for Wisconsin public schools.

J. BARKLEY ROSSER, JR., Ph.D. Wisconsin (Madison) 1976. Essays related to spatial discontinuities in land values at the urban-rural margin.

KIRK W. RUBIDA, Ph.D. Colorado 1977. An econometric model of Colorado economy.

RONALD RUDOLF, Ph.D. Rutgers 1976. Physician care and government programs: An analysis of the distribution of health and health services in New York City and the nation.

DAVID RUTHENBERG, Ph.D. Wayne State 1976. Determination of interstate variation in state expenditures

JAMES A SACKEY, Ph.D. Western Ontario 1977. An economic analysis of the production of hospital services in Ghana.

JOSE M. SANCHEZ-MOLINERO, Ph.D. California (Santa Barbara) 1976. Discrimination as an allocation mechanism in the labor market· Theory and empirical evidence.

FARHAT S. SHERNANNA, Ph D Oklahoma State 1976. A comparative evaluation of public housing programs in the Libyan Arab Republic.

IRA SILVER, Ph D City (New York) 1977. Regional input-output forecasting· A systems approach.

MARVIN M. SMITH, Ph D Cornell 1977 Essays on the economics of discrimination in the labor market

ALI A SOLIMAN, Ph.D Iowa 1975. Reimbursement and hospital pricing and utilization

RICHARD G STEVIE, Ph.D. Cincinnati 1977 The economics of metropolitan water supply.

CHARLES T. STREIN, Ph D Illinois 1977 The university as a nonprofit firm. Administrative expense bias as a source of inefficiency

ROBERT J STROM, Ph D Cincinnati 1976 The effect of input characteristics on the conceptual content of high school courses

JOSEPH M SULOK, JR , Ph D Virginia 1976 A theoretical and empirical analysis of the demand for education at community colleges

MEHMET TAHIR, Ph.D Northeastern 1976. An economic analysis of the utilization of pediatric nurse practitioners

VINOD THOMAS, Ph.D. Chicago 1977. The welfare cost of pollution control with spatial alternatives.

THOMAS H. TIDRICK, Ph.D. Georgia State 1977. Educational production function with alternative technologies.

ROBERT P. TROST, Ph.D. Florida 1977. Estimation of a housing expenditure model with interdependent choices about owning or renting.

JEFFREY G. TOWLE, Ph.D. Michigan 1977. A new approach to classifying media vehicle audiences.

ALAN P. WAGNER, Ph.D. Illinois 1977. The impact of federal-student BEOG programs on low income household expenditure for higher education.

GEORGE B. WALL, Ph.D. North Carolina State 1977. Shared use of public school facilities as a capital investment decision policy.

LYNN B. WARE, Ph.D. Boston College 1976. The demand for general relief assistance· A conditional probability analysis of employment and job tenure transactions.

JAMES O. WATSON, Ph D. Purdue 1976. Estimated rates of return to investments in public and proprietary postsecondary vocational education and the desirability of subsidizing proprietary institutions.

JOSEPH C WEBER, Ph D. Colorado State 1976. An interindustry economic analysis of northwestern Colorado.

ROBERT S WENGER, Ph.D. Nebraska 1976. A forecast of the household final demand vector of the Nebraska input-output tables.

GLENN D WESTLEY, Ph D. Pennsylvania 1976. The optimal location of urban radial lines

JOHN R C. WHEELER, Ph.D. Michigan 1976. Economic efficiency and optimal scale in ambulatory medical care production

MARTIN WILLIAMS, State University of New York (Binghamton) 1977. The determinants of modal choice in urban travel· A case study.

FLOYD S. WILSON, Ph D. Minnesota 1976 Regression and choice models to estimate the enrollment effects of campus closure.

EDWIN A. WOOD, Ph.D Florida 1976. Quantitative factors for instructional resources in the state university system of Florida

LIH YING H. YOUNG, Ph.D. City (New York) 1977 Household demand for leisure activities· Reading and TV watching

PHILLIP YOUNG, Ph.D. New York 1977 An analysis of the economic returns of formal schooling for Puerto Rican men in New York City.

ANNOUNCEMENT

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1978 issue of the *Review* will carry the seventy-fifth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1978. This announcement is an invitation to send us information for the preparation of the list. This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1978 issue the names of those who will have been awarded the doctoral degree since June 1977 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3 × 5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

JEL Classification No. ______

Name: LAST NAME IN CAPS· First Name, Initial

Institution Granting Degree: ______

Degree Conferred (Ph.D. or D.B.A.) ______ Year ______

Dissertation Title. ______

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

The following Statement of Ownership, Management, and Circulation is provided in accordance with the quirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and pı lished by the American Economic Association, a nonprofit educational organization, located at 1313 21st Av nue So., Nashville, Davidson County, Tennessee 37212. The Managing Editor is Professor George H. Bor *American Economic Review*, Brown University, Robinson Hall, Providence, R.I. 02912. During the preceding months the average number of copies printed for each issue was 27,500, the average paid circulation, 24,232; average free distribution, 354; the average number of copies distributed, 24,586. Corresponding figures for last issue before filing: 27,500 total number copies printed; 25,268 total paid circulation; 354 free copies dis buted; 25,622 total distribution.